EUROPEAN LANGUAGE EQUALITY

D2.15

Technology Deep Dive – Text Analytics, Text and Data Mining, NLU

Authors	Jose Manuel Gomez-Perez, Andres Garcia-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiļjevs, Teresa Lynn

Dissemination level	Public
Date	28-02-2022

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	D2.15
Deliverable title	Technology Deep Dive – Text Analytics, Text and Data Mining, Natural Language Understanding
Type	Report
Number of pages	64
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP2: European Language Equality – The Future Situation in 2030
Task	Task 2.3 Science – Technology – Society: Language Technology in 2030
Authors	Jose Manuel Gomez-Perez, Andres Garcia-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiļjevs, Teresa Lynn
Reviewers	Itziar Aldabe, Georg Rehm
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

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 Tables

List of Acronyms

ABSA	Aspect-Based Sentiment Analysis
AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech Recognition
B2B	Business to Business
B2C	Business to Customer
CALL	Computer Assisted Language Learning
CLARIN	Common Language Resources and Technology Infrastructure
CLIP	Contrastive Language–Image Pre-training
CNN	Convolutional Neural Network
DL	Deep Learning
DLE	Digital Language Equality
DPP	Data Protection and Privacy
EHR	Electronic Health Record
EL	Entity Linking
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRC	European Language Resource Coordination
EU	European Union
FAQ	Frequently Asked Questions
GDPR	General Data Protection Regulation
HITL	Human-in-the-loop
HPC	High-Performance Computing
ICT	Information Communication Technology
IE	Information Extraction
IID	Independent and Identically Distributed
ISRL	Implicit SRL
IT	Information Technology
LM	Language Model
LR	Language Resources/Resources
LT	Language Technology/Technologies
MAPA	Multilingual Anonymisation for Public Administrations
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NDA	Non-Disclosure Agreement
NED	Named Entity Disambiguation

NER	Named Entity Recognition
NLG	Natural Language Generation
NLP	Natural Language Processing
NLTP	National Language Technology Platform
NLU	Natural Language Understanding
OIE	Open Information Extraction
PII	Personal identifiable information
POS	Part-of-Speech
QA	Question Answering
R&D	Research and Development
RE	Relation Extraction
SOTA	State-of-the-Art
SRL	Semantic Role Labelling
TA	Text Analysis
UK	United Kingdom
VA	Virtual Assistant
VSA	Virtual Scientific Agent
W3C	World Wide Web Consortium
WHO	World Health Organization
WSD	Word Sense Disambiguation

Abstract

Text analytics and natural language understanding (NLU) deal with extracting meaningful information and insights from text as well as enabling machines to understand such content in depth, similar to how a human would read a document. These tools have been on the market for several years and have successfully found applications in many sectors including health, education, legal, security, defense, insurance, and finance to name but a few. However, existing text analytics and NLU services do not cover all languages equally.

The market offer around these technologies tends to gather around those languages that cover a larger segment of the population, maximizing the return on investment. As a consequence, there is a risk of discrimination in terms of the coverage provided to European languages with a lower number of speakers despite current efforts to ameliorate this situation. To reduce the coverage gap across languages both in the market and in society, technical, regulatory, and societal advances are required that increase access to text analytics and NLU technologies regardless of the specific European language and territory. Among others, the creation of large datasets and benchmarks across the different languages and verticals, as well as a policy of incentives to stimulate the service offering in underrepresented languages will be key.

Neural language models are a key data-driven emergent technology in text analytics and NLU, with the potential to revolutionize the offer of text understanding functionalities and to increase the coverage of such tools for less widely spoken languages. Language models have proven to be very useful to solve tasks like key phrase extraction, named entity recognition, relation extraction, classification, and sentiment analysis and have made impressive progress on tasks that were considered experimental and not ready for the market yet, such as question answering or abstractive summarisation. Building language models for less widely spoken languages is therefore a strategic step towards digital language equality.

While training neural language models is a self-supervised process that does not require annotated data, fine tuning such models to address specific tasks does require annotated datasets. The availability of textual resources across the different European languages is therefore an important factor to leverage the full potential of language models. Unfortunately, not all languages are equally provisioned with such resources. Moreover, annotated data in multiple sectors and industry use cases is scarce, hampering the use of language models and in general of data-driven approaches to text processing. Textual resources, annotated data, and techniques that work well in low resource scenarios (self, weakly or semi-supervised) are important assets to build and apply effective language models.

In this document, we present a comprehensive overview of text analytics and NLU tools under the perspective of digital language equality in Europe. We focus both on the research that is currently being undertaken in foundational methods and techniques related to these technologies as well as gaps that need to be addressed in order to offer improved text analytics and NLU support in the market across languages. Our analysis ends with a succinct list of eight recommendations and guidelines that addresses central topics for text analytics and NLU. Such topics include among others the role of language equality for social good, the balance between commercial interests and equal opportunities for society, and incentives to language equality, as well as key technologies like neural language models and the availability of cross-lingual, cross-modal, and cross-sector datasets and benchmarks.

1 Introduction

Text analytics tools have been in the market for several years and have proved useful to extract meaningful information and insights from documents, web pages and social media

feeds, among other text sources. Text analysis processes are designed to gain knowledge and support strategic decision making that leverages the information contained in the text. Typically, such a process starts by extracting relevant data from text that later is used in analytics engines to derive additional insights. Nowadays text analysts have a wide range of accurate features available to them to help recognize and explore patterns, while interacting with large document collections.

Text analysis is an interdisciplinary enterprise involving computer science techniques from machine learning, information retrieval, and particularly natural language processing. Natural language processing is concerned with the interactions between computers and human (natural) languages, and, in particular, with programming computers to fruitfully process large natural language corpora. Challenges in natural language processing frequently involve natural language understanding, natural language generation, connecting language and machine perception, dialog systems, or some combination thereof.

Recent breakthroughs in deep learning have made impressive progress in natural language processing. Neural language models like BERT and GPT-3, to name some of the most widely-used, are able to infer linguistic knowledge from large collections of text that then can be transferred to deal effectively with natural language processing tasks without requiring too much additional effort. Neural language models have had a positive impact in key features of text analytics and natural language understanding, such as syntactic and semantic analysis, entity recognition and relation extraction, text classification, sentiment analysis, machine reading comprehension, text generation, conversational AI, summarisation, and translation, among others.

The success of machine and deep learning has caused a noticeable shift from knowledge-based and human-engineered methods to data-driven architectures in text processing. The text analytics industry have embraced this technology and hybrid tools are incipiently emerging nowadays, combining or replacing robust rule-based systems that have been the norm in the market until now with machine learning methods. Nevertheless, despite all the hype about data-driven approaches to text processing and particularly transformer language models like BERT (Devlin et al., 2019), which might lead to thinking that everything is already solved in text analysis and language understanding, there are still many gaps that need to be addressed to make them fully operational and to benefit all European Languages. Especially relevant is the fact that data-driven approaches require large amounts of data to be trained.

Language models have lessened the requirement of labelled data to address downstream tasks, yet the need for such data has not disappeared. Beyond general purpose datasets, labelled data is scarce, labor intensive and therefore expensive to generate. Labelled data is one of the major burdens to leverage data-driven approaches in business applications and is also problematic for under-resourced or minority languages for which such data does not exist and there is little interest from technology providers to produce it. Moreover, neural language models work as black boxes that are hard to interpret. This lack of transparency makes it difficult to build trust between human users and system decisions. Lack of explanation abilities is a major obstacle to bring such technology in domains where regulation demands systems to justify every decision. Furthermore, language models face ethical challenges including gender and racial biases that are learnt from biases present in the data the models are trained on, thus perpetuating social stereotypes.

While the progress made in the last years is undeniably impressive, we are still far from having perfect text analytics and natural language understanding tools that provide appropriate coverage to all European languages, particularly to minority and regional languages. Thus, one of the main goals of this document is to define a 10-year research roadmap that helps the European text analytics industry and research community to address the shortcomings and builds upon the strengths of current text analytics and natural language understanding tools. We call for human-centric text analysis where people's knowledge, emotions and needs are put at the center of design and learning process of the next generation of text

analytics tools. Other topics in the research agenda are hybrid approaches to natural language processing, combining existing rule-based and data-driven systems, multilingualism in text analytics, multimodal analysis of information, and a new generation of benchmarks for natural language processing tools.

The remainder of this document is structured as follows: The scope of the document is described in Section 2. The main text analytics components currently supported by text analytics tools, including an analysis of language coverage across European languages, are presented in Section 3. Next, Section 4 is devoted to describing the state-of-the-art (SOTA) in research areas related to text analysis and natural language understanding. Based on such analysis of the state-of-the-art, we identify key issues, gaps and challenges in Section 5. The impact of text analysis in society and its contribution to digital language equality is discussed in Section 6. The research roadmap is split into Section 7 and Section 8, where the former focuses on the breakthroughs needed to advance the state-of-the-art and fill the gaps in text analysis and the latter focuses on the technology visions and development goals. Then, we look forward into deep natural language understanding and its contribution to general AI in Section 9. Finally, we present a summary and conclusions of the document in Section 10.

2 Scope of this Deep Dive

This document aims at collecting, analyzing and consolidating the views of European research and industrial stakeholders on the progress towards digital language equality in core text analytics and natural language understanding (NLU) technologies, innovations, and impact on society ten years from now. To better understand how these technologies are currently being made available to end users, stakeholders and society, we adopt a multidimensional approach where both a market and research perspective are considered, as well as the key domains and applications related to text analytics and NLU.

We look at the current service and tool offering of the main text analytics and NLU providers in the European market. This analysis also includes recent findings in related research areas, such as natural language processing and understanding, machine learning, and information retrieval, where language understanding tasks that not long ago were subject of study in research laboratories are now part of the text analytics market. This is as a result of recent breakthroughs in deep learning, structured knowledge graphs and their applications.

Conventional text analytics services available in the market include syntactic analysis, extractive summarisation, key phrase extraction, entity detection and linking, relation extraction, sentiment analysis, extraction of personal identifiable information, language detection, text classification, categorization, and topic modeling, to name but a few. Also, conversational AI services and tools, including chatbots and virtual agents, are frequently offered under the umbrella of text analytics. More recent additions to the text analytics catalogue are machine reading comprehension services based on tasks such as extractive question answering, which are usually marketed as part of both virtual agents and intelligent search engines to provide exact answers to user questions.

In addition to general-purpose text analytics, we also consider in this document specific domains where text analytics technologies are particularly important. For example, there is a significant number of specific text analytics tools focused on Health, including functionalities such as extraction of medical entities, clinical attributes, and relations, as well as entity linking against medical vocabularies. Other use cases for text analytics tools include customer experience, employee experience, brand management, recruiting, or contract analysis, to name a few. However, an exhaustive account of each sector and use case, and their relevance for text analytics is out of the scope of this document.

Nowadays, text analytics tools and services are available for widely spoken languages or

otherwise strategic languages where the market is big enough for companies to make a profit. Unfortunately, other languages may be less attractive from a business point of view and consequently they are not equally covered by the current offer of text analytics tools. Throughout this document, language coverage is addressed as another key dimension for the analysis of text analytics and NLU tools for digital language equality.

We include recent research breakthroughs associated with the text analytics services mentioned above. Many applications of text analytics can be effectively solved using classical machine learning algorithms, like support vector machines, logistic regression or conditional random fields, as well as rule-based systems, especially when there is little or no training data available. Actually, it is good practice to always use the most simple approach possible to solve a language problem. However, more sophisticated approaches are needed as we transit towards scenarios involving a deeper understanding of text in order to solve increasingly complex tasks like abstractive summarization, reading comprehension, recognizing textual entailment, or stance detection. Therefore, this document makes special emphasis on deep learning architectures, like transformer language models, and their extensions.

Of particular interest for language equality are different means to deal with data scarcity for low-resource languages. Self-supervised, weakly supervised, semi-supervised, or distantly supervised machine learning techniques reduce the overall dependence on labeled data, but even with such approaches, there is a need for both sufficient labeled data to evaluate system performance and typically much larger collections of unlabeled data to support the very data-hungry machine learning techniques. Also in this direction, we include a discussion on hybrid approaches where knowledge graphs and deep learning are used jointly in an effort to produce more robust, generalisable, and explainable tools. In addition, we consider research addressing multilingual and cross-lingual scenarios.

Another important area of research that we touch upon in this document deals with leveraging other modalities of information in addition to text. For example, images, figures and diagrams in the scholarly and health domain can provide additional context for document analysis. Similarly, in customer experience systems, speech data is used in conjunction with text to gather signals from different channels. Also the other way around, text in captions or accompanying posts in social media, e. g., Twitter, can be used to help with image processing and classification.

Finally, all the above-mentioned aspects are taken in consideration from the perspective of their combined impact in society. In doing so, it is the objective of this document to provide a series of recommendations as to how to address the current limitations of text analytics and natural language understanding technologies and their contribution to digital language equality.

3 Text Analytics: Main Components

The goal of text analytics is to discover novel and interesting information from documents and text collections that is useful for further analysis or strategic decision making. Text analytics tools can extract structured data from unstructured text, classify documents in one or more classes, label documents with categories from taxonomies, and assign a sentiment or emotion to text excerpts, among other functionalities. Such structured information, including data, classes, categories, labels, sentiments, and emotions, is then used to fuel analytic tools and find patterns, trends, and insights to improve tasks such as search and recommendation and, in general, supporting through automation the accomplishment of any task involving large amounts of text processing.

Text analytics tools support a wide range of functionalities to process and leverage text. Most of these functionalities can be broadly categorized into syntax analysis, information

extraction (e. g., key phrases, entities, relations, and personal identifiable information), text classification, sentiment and emotion analysis, and conversational AI functionalities. Recently, Question Answering, a functionality requiring machine reading comprehension, has made the transition from research labs to production systems. Below we describe some of the most frequent functionalities supported by text analytics tools. Part of this analysis is based on input from market studies in Language Technologies, such as Gartner Magic Quadrant for Insight Engines (Emmott and Mullen, 2021) and The Forrester Wave: AI-Based Text Analytics Platforms 2020 (Evelson et al., 2020).

Syntax analysis Syntactic analysis refers to the process of linguistically parsing text into a format that is useful for downstream tasks. It involves taking raw text and breaking it into a series of sentences or words, and for each word identifying the lemma (dictionary entry), part of speech (e. g., noun, verb) or inflectional information (e. g., plural form). Finally, the syntax or grammatical structure is specified through identifying the relationship between the words (e. g., subject, clause, nominal modifier, etc.).

Key phrase extraction The process of identifying key phrases in a text or corpus. A key phrase is a relevant part of the text. Different key phrase types are usually offered: main phrases, main lemmas, main concepts, or relevant topics. For example, in the text "The hotel was amazing and the staff were incredible.", key phrase extraction might return the main topics: "hotel" and "incredible staff".

Entity extraction and linking Entity extraction, also known as entity name extraction or named entity recognition, is a technique that identifies key elements from text, then classifies them into predefined categories. Entity linking disambiguates the identity of entities according to a pre-existing resource, such as a knowledge base. For example, in the sentence "We went to Seville last week.", the entity extraction process would identify "Seville" as a location entity and the linking process would link it to more information in its Wikipedia entry.

Relation extraction The process of identifying and classifying relations between entities in text and/or data. These relations can be expressed as a verb plus the text elements that are in a semantic relation with it. For example, given the text "John sent a letter to Mary.", the verb "sent" is related to "John" as the subject,"a letter" as the object, and "to Mary" as the target.

Summarisation The process of reducing one or more textual documents to create a summary that retains the most important points of the original document(s). The most common approach to summarize text is to extract sentences that collectively represent the most relevant information within a document. Recently research has moved towards abstractive summarization where the goal is to generate a summary by rephrasing the original text. This is useful, for example, in medical or scientific research.

Personal identifiable information (PII) detection Detection of entities in a text that contain personally identifiable information (PII), or PII entities. A PII entity is a textual reference to personal data that could be used to identify an individual, such as an address, bank account number, or phone number. PII underlies the process of anonymisation of text.

Sentiment and emotion analysis The process of identifying and categorizing opinions expressed in a piece of text, especially to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral.

Text Classification or text categorization The process of grouping documents into classes or categories. An example would be the classification of customer reviews into comments that may require action, removal (harmful content) or consideration (suggestions for improvement).

Language detection The process of guessing which natural language a text or text segment is written in. This is a fundamental task when dealing with big data that is crawled from a multilingual source (e. g., the web).

Chat bots or virtual agents A chatbot is an interactive computer program that uses artificial intelligence (AI) and natural language processing (NLP) to understand user questions and automate responses to them, simulating human conversation. In order to understand the user's current goal, the system must leverage its intent detector to classify the user's utterance into one of several predefined intents.

Question Answering The process where computer systems answer questions posed by users in the form of natural language. A common approach is extractive question answering, which is the task of extracting an answer for a question from a document or collection of documents (e. g., "What are the current banking fees?"). At a higher level, there is open-domain question answering, which aims to answer a question based on large-scale unstructured documents.

The challenges involved in the different tasks in natural language processing and understanding have different levels of complexity and as a result the solution to each of such challenges is in different degrees of progress. For example, natural language generation is one of such challenges, where recent advances like GPT-3[1] are currently producing new achievements. Therefore, in addition to functionalities that are already available in the market, there are others on which the research community is currently working.

Some advanced functionalities involve reasoning capabilities such as **multi-hop question answering** where systems need to gather information from various parts of the text to answer a question, and **textual entailment**, where the goal is to determine whether a hypothesis is true, false, or undetermined given a premise. Moreover, with the advent of **generative models** like GPT-3 new opportunities arise to address hard problems involving text generation. For example, **abstractive text summarisation**, where the system generates a summary of a text rather than extracting relevant excerpts, or **data to text generation**, where the goal is to generate text descriptions out of data contained in tables or json documents. Furthermore, researchers are working on **stance detection**, a functionality that has proven useful to deal with misinformation and fake news. See for example the work of ALDayel and Magdy (2021) for a survey on stance detection in social media. With stance detection, a system can identify whether a fact-checked claim supports or refute another claim made, e. g., in a news article. Claims refuted by fact-checked claims can be regarded as low credibility statements.

3.1 Custom Text Analytics Using Machine Learning

Recently, commercial text analytics providers have started supporting the customization of functionalities. For example, users can define the classification classes, entity and relation types, or sentiment scores. This customization is possible thanks to supervised learning where a machine learning model learns from user-generated examples. The user input is limited to providing the examples, while the text analytics tool handles all the complexity of

[1] Actually, GPT-3 is being marketed as a core infrastructure to fuel the next generation of applications involving language generation (see https://openai.com/blog/gpt-3-apps/)

the machine learning process, including the learning algorithm, parameter tuning or model pre-training. Thus, end users do not need a strong machine learning background to customize their own services. However, some basic knowledge is required to understand how the trained models are evaluated and how to generate a balanced set of examples.

The most common customisable text analytics services are classification and entity extraction. Nevertheless, providers typically offer support for sentiment analysis and relation extraction, too. To customise a text classifier users need to provide examples of text labeled with classes, for entity extraction the text is labeled with entity types, for relation extraction relations between entities are indicated, and for sentiment analysis documents are labeled with a sentiment score.

3.2 Language Support

To study the language support of existing text analytic technologies and natural language understanding tools, we look in two main directions. The first source of interest for this analysis is the catalogue of services of global technology providers, which provide us with a notion of what is being currently made available and marketed to the public. Then, we look into European initiatives that offer repositories of language resources and tools. We base our analysis on the catalogue of the European Language Grid (ELG). At the time of writing, the ELG catalogue holds more than 11,500 metadata records, encompassing both data and tools/services, covering almost all European languages – both official and regional/minority. The ELG platform was populated with more than 6,000 additional language resources that were identified and documented by language informants in the ELE consortium and harvests many major EU LR/LT repositories such as CLARIN[2] and ELRC-SHARE.[3] Our goal was not to provide an exhaustive account, for which such figures could be complemented with additional information from other European infrastructure like the ones above mentioned, but an indicator of the current language support at the European level.

In the case of commercial services, we choose key players in text analytics market reports such as Gartner Magic Quadrant for Insight Engines and The Forrester Wave: AI-Based Text Analytics Platforms 2020. A mandatory requirement for providers to be included in this study is that the documentation of the services is publicly available. We study services and languages supported by Azure Text Analytics, IBM Watson, Expert.ai and SAS Visual Text Analytics. In addition, we include in this list other recognized technology providers such as Amazon Comprehend and Google Natural Language API.

To simplify the analysis of the language support we use the following groups:

- A – Official EU Languages (24): Bulgarian, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Irish, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Slovak, Slovenian, Spanish, and Swedish
- B – Other EU languages; languages from EU candidate countries and Free Trade Partners (11): Albanian, Basque, Catalan, Galician, Icelandic, Norwegian, Scottish Gaelic, Welsh, Serbian, Turkish, Ukrainian
- C – Languages spoken by EU immigrants; languages of important trade and political partners (18): Afrikaans, Arabic, Berber, Cebuano, Chinese, Hebrew, Hindi/Urdu, Indonesian, Japanese, Korean, Kurdish, Latin, Malay, Pashto, Persian (Farsi), Russian, Tamil, Vietnamese

In Table 1 we report the language support offered by global text analytics providers across different text analytics services or functionalities. A small set of services including Entity

[2] https://www.clarin.eu
[3] https://elrc-share.eu

Extraction, Key Phrase Extraction, and Syntax analysis have a large coverage, above 80%, of EU official languages in category A. Nevertheless, the support of the languages in category A provided by the rest of the services is poorer, ranging from 20% to 45%. The situation of other EU languages in category B is actually the worst: the language support of the functional services is scarce or directly non-existent. Languages in category C also have low coverage across all functional services.

	Category		
Functional Service	**A**	**B**	**C**
Entity Extraction	22	3	11
Key Phrase Extraction	21	3	9
Syntax Analysis	19	2	9
QA	11		3
Sentiment Analysis	9	2	9
Chatbot	8		4
Classification	7		4
Summarization	6		3
Relation Extraction	6		4
PII	6		3

Table 1: Language support of text analytics services by global technology providers.

In addition, in Table 2 we report on the language support of custom text analytics services. Custom Entity Extraction has almost perfect support of languages across all the categories. However, custom classification, custom sentiment analysis, and custom relation have a language coverage similar to off-the-shelf text analytics services, covering less than half of the official languages in category A and C, and almost none of the languages in category B.

	Category		
Functional Service	**A**	**B**	**C**
Custom Entity Extraction	23	10	17
Custom Classification	11	1	7
Custom Sentiment Analysis	11	1	7
Custom Relation Extraction	7		4

Table 2: Language support of custom text analytic services by global technology providers.

To complement this analysis we include the ELG catalogue of functional services. ELG aims at being the primary platform and marketplace of language technologies in Europe. The ELG catalogue of services is much more fine grained. It includes services for specific natural language processing tasks, such as date detection or numerical annotation, which industry providers often bundle into broader services.

In Table 3 we report the number of languages supported by ELG functional services in each language category.[4] There is a small group of services, at the top of the table that represent syntax analysis – language identification, tokenisation, lemmatisation, morphological analysis, part-of-speech tagging, and dependency parsing – tools for which are available for all or nearly all languages in category A. Nevertheless, the language support of such services drops to 63% of languages in category B, and 72% in category C. Named entity recognition is

[4] This is a snapshot of the ELG release 2 from November 2021

indicated as having moderate support across all language categories reaching 66% for category A, 54% for category B and 61% for category C. Following on from there, the language support of main text analytics services such as keyword extraction, sentiment analysis, summarisation, and entity linking is poorer or non-existent in every language category.

	Category		
Text Analytics Service	**A**	**B**	**C**
Dependency Parsing	24	7	13
Lemmatization	24	7	13
Morphological analyser	24	7	13
Part-of-Speech Tagging	24	7	13
Tokenization	24	7	13
Language Identification	22	6	14
Named Entity Recognition	16	5	11
Keyword Extraction	10	3	11
Sentiment Analysis	9		4
Semantic Annotation	7		5
Summarization	7		5
NER Disambiguation	4		
Sentence Splitting	4		
Textual Entailment	3		
Date Detection	2		
Entity Linking	2		
Text Classification	2		
Discourse Parsing	1		
Information Extraction	1		
Measurement Annotation	1		
Measurement Normalisation	1		
Negation Detection	1		
Noun Phrase Extraction	1		
Number Annotation	1		
Number Normalisation	1		
Opinion Mining	1		
Parsing	1		
Text Extraction	1		

Table 3: Number of distinct languages covered by functional services in ELG

Both sources of information, global text analytics providers and ELG, show us that official EU languages are covered by a subset of text analytics services including syntax analysis, key phrase extraction, and entity extraction. However, the numbers reported in the tables above are absolute counts on the presence of a tool, but it is not possible to assess its technical readiness/fit-for-purpose at this level. Moreover, only a small fraction of languages in this category are supported by the rest of the services. For other EU languages in category B, global players offer scarce support or no support at all, and for languages in category C the support is also low. In ELG the picture changes somewhat for languages in category B since the number of supported languages increases for some of the functional services. Nevertheless, the overall support of languages in category B and C is still low.

The differences between the three tables point at different priorities in terms of the development and offering of text analytics functionalities followed by commercial technology providers and publicly funded initiatives like ELG. Figures seem to indicate that the former

plan their offering in terms of the volume of the potential market each specific language can lead to. On the other hand, the latter seem to be guided not only by the principles of offer and demand, but also by the underlying aim to make all technologies accessible by all European citizens, equally.

4 Text Analytics: Current State of the Art

In this section we will analyze state-of-the-art technologies in Text Analysis (TA) and Natural Language Understanding (NLU). TA is an AI technology that uses Natural Language Processing (NLP) tools and techniques to extract relevant information from large amounts of unstructured text. It is a key enabling technology that allows building data-driven approaches to manage and interpret textual content, and therefore developing applications that are able to carry out various types of user –or business– driven analyses on written text. Natural Language Understanding (NLU) is a subset of NLP, whose aim is to understand human language text on a semantic level, and has many applications such as Question Answering, Machine Translation or Chatbots, to name a few.

We will start by describing core technologies that allow building TA solutions. We will also put a special focus on neural language models, which are particularly useful for developing TA systems for tasks and languages where manually annotated examples are scarce. Finally, we will also point out specific TA applications in different domains.

4.1 Text Analysis and Natural Language Understanding

The main goal of TA is to generate structured data out of free text content by identifying facts, relationships and entities that are buried in the textual data. 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. Once the information that is implicitly conveyed in text is made explicit, it can be stored in a structured way and further processed by user and business analytic tools. Moreover, in our increasingly multilingual world this information should be processed in multiple languages to allow for a cross-lingual and interoperable semantic interpretation. Ideally, this processing is robust enough to provide the same accurate results in multiple application domains and textual genres.

To make the problem more manageable, TA is addressed in several tasks that are typically performed in order to preprocess the text to extract relevant information. The most common tasks currently available in state-of-the-art NLP tools and pipelines include part-of-speech (POS) tagging, 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. **Word Sense Disambiguation** (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 over any type of word, such as nouns, verbs or adjectives, as well as over 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 for a number of languages with current neural language models (Akbik et al., 2019; Devlin et al., 2019), POS tagger accuracy still degrades significantly when applied on out of domain data (Manning, 2011). Closely

related to POS tagging is lemmatization (obtaining the canonical word or lemma from a given word form), which has traditionally been considered to be crucial for POS tagging.

For a text analysis system to be able to recognize, classify and link every mention of a specific named entity in a document, several tasks are considered, namely, NER, **Named Entity Disambiguation 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 recognize a named entity, classify it as a type (e.g, person, location, organization, etc.), and resolve 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).

Semantic Role Labelling 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.

Information Extraction (IE) aims to derive structured information 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., 2015), RNNs (Nguyen and Grishman, 2016) and Transformers (Yang et al., 2019) to address this task. 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 (e. g., cross-lingual transfer).

Another important task within IE is **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 pre-trained 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., 2021b) freezes the parameters of a pretrained LM and utilizes Adapters,[5] to leverage factual knowledge from Wikipedia as well as syntactic information in the form of dependency parsing.

Once the main events are identified, **Temporal Processing** aims to capture and structure Temporal Information. This consists of 1) identifying and normalizing any temporal expression and event in the text and 2) establishing the temporal order in which the events oc-

[5] Adapters, originally proposed by Houlsby et al. (2019) have been introduced as an alternative lightweight fine-tuning strategy that achieves on-par performance to full fine-tuning on most tasks. They consist of a small set of additional newly initialized weights at every layer of the transformer. Created by Pfeiffer et al. (2020), https://adapterhub.ml offers a framework and repository for pretrained adapter modules.

curred, as defined by the TempEval3 shared evaluation task (UzZaman et al., 2013).

To summarize, 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, Summarization 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 Aspect Based Sentiment Analysis (ABSA), with the opinions and expressed polarity (positivity or negativity) referring to each event and its participants (Vossen et al., 2016). Aspect Based Sentiment Analysis (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).

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. When there is not enough data manually annotated by linguists for a semantic task in a given language, major obstacles arise when training supervised models. In most cases, manually annotating text for every single specific need is generally inefficiently slow and, in most cases, not affordable in terms of human resources and economic costs. Even when manually annotated resources are available, a common problem that researchers face is that texts need to be accurately analyzed at many distinct levels for a full understanding. Furthermore, each of these levels are affected by ambiguous expressions that cannot be interpreted in isolation.

4.2 Neural Language Models

TA is undergoing a paradigm shift with the rise of *neural language models*[6] that are trained on broad data at scale and are adaptable to a wide range of monolingual and multilingual downstream tasks (Devlin et al., 2019; 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).

One of the most pressing problems in TA is the scarcity of manually annotated examples in real world applications, particularly when there is a domain and language shift. In such circumstances, traditional machine learning 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.

The aforementioned methods are examples of *transfer learning*, whose main idea is to take the "knowledge" learned from one task (e. g., predict the next word given the previous words) and apply it to another task (e. g., information extraction). This way, models are able

[6] Also known as Pre-trained Language Models (Han et al., 2021)

to leverage previous learning, and avoid starting the training process from scratch. Within deep learning, pre-training is the dominant approach to *transfer learning*: the recipe is to *pre-train* a deep transformer model on large amounts of unlabelled data and then reuse this pre-trained language model by *fine-tuning* it on small amounts of (usually annotated) task-specific data. This means that, even for a traditionally complex task such as Coreference Resolution (Pradhan et al., 2012), current transfer learning approaches based on pre-trained language models obtain state-of-the-art results, even without requiring extra linguistic annotations. Nevertheless, annotated data is still required to evaluate the models in downstream tasks.

4.3 Applications

One of the main applications of TA systems is allow humans to interact with computers using natural language. Examples of these systems are **chatbots** or **virtual agents**, which engage users to have conversations with them. Popular systems include Siri,[7] Google Assistant,[8] and Amazon Alexa,[9] among others. A related application are the so called **task-oriented dialogue systems**, which maintain a conversation with users and help them to perform a concrete task, such as booking a table at a restaurant, calling someone or checking the weather forecast. Virtually all these systems they include a NLU module, whose aim is to analyze user utterances to input the intent and extract relevant information in form of slots or concepts. Chatbots and dialogue systems also include a Natural Language Generation (NLG) module, whose objective is to generate the responses to the user.

Another important area of application are **interactive question answering systems**, systems that allow users to express their information need using natural language, and are able to answer the queries posed by users by analyzing a large quantity of documents. Usually, question answering (QA) systems are classified into extractive and abstractive. The former uses NLU techniques to understand the query and documents, and return selected excerpts from the document as the final answer. Abstractive QA systems use NLG to generate the final response to the user, based on the facts that are identified in the documents. In both cases, the core technology is commonly based on pre-trained language models (Section 4.2), with additional mechanisms to represent the context (Huang et al., 2019). Lately, *conversational agents* have emerged, as an hybrid between QA systems and chatbots. Conversational agents meet user information needs by having conversations with them, often by emulating the personality of a human (Zhang et al., 2018). The Alexa prize,[10] 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).

Generation of new text (NLG) is one of the main applications of TA systems (Gehrmann et al., 2021). Example applications that generate new texts from existing (usually human-written) text include **machine translation from** one language to another, **summarisation**, **simplification**, text correction, paraphrases generation, question generation, etc. Nowadays, NLG is often achieved by means of deep learning neural architectures (Li et al., 2021). 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 pre-trained language models. This way, the models are able to encode massive linguistic and world knowledge accurately

[7] https://www.apple.com/es/siri/
[8] https://assistant.google.com
[9] https://www.amazon.com
[10] https://developer.amazon.com/alexaprize

and express in human language fluently, both of which are critical abilities to fulfill text generation tasks. For text generation tasks, some of the pre-trained language models utilize the standard Transformer architecture following the basic encoder-decoder framework, while others apply a decoder-only Transformer. 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.

5 Text Analytics: Main Gaps

During the past decade, the ecosystem related to text analytics, text and data mining, and natural language understanding has changed and improved dramatically. This is due amongst other things, to advances in the area of Deep Learning. Nonetheless, issues, gaps, and challenges still exist. In this section on gaps, we will break these down into 7 main areas:

1. Data
2. Legal
3. NLU system limitations
4. Benchmarking
5. Investment protection and interoperability
6. Conformance
7. Consumer-grade tool support for domain experts

5.1 Data

The availability of suitable data for use in both training and evaluating today's state-of-the-art NLP tools is crucial. Unfortunately, the state-of-affairs related to language data for text analytics suffers from a number of shortcomings.

The type of data required for TA tools can vary according to the task at hand. For example, when building large transformer-based language models, current systems can be built upon raw (unlabelled) text. Collections of digital text from various sources such as Wikipedia, websites, books, etc., can be combined to form such a suitable raw corpus. However, the main text analytic tools discussed in Section 3 (i. e., more sophisticated tasks such as named entity recognition, syntactic parsing, sentiment analysis, etc.) require data to be labelled in such a way that the model can learn and induce patterns, therefore enabling label predictions to be made on new or previously 'unseen' data.

Labelling data can be a **time-intensive task** that often requires skilled domain expertise, both of which are costly overheads for both the research and industry communities. **Data coverage** is also an important consideration. While general language data may be useful for developing a language model, domain-specific language data (e. g., medical, legal, user-generated content, etc.) may be needed to ensure sufficient coverage of certain terminology and phrasing. Likewise, **language coverage** is a concerning issue as the majority of datasets being produced that are relevant to Europe are based on the major languages such as English, German, Spanish and French. Within all of these datasets, **quality** is also important. Quality in terms of having reliable content (i. e., no fake news), balanced content (e. g., no bias) and clean content (i. e., non-toxic/hate-speech). Machine learning models are notoriously sensitive to bias and noise within datasets. Thus, biased data will lead to biased predictions. There is a clear need, therefore, for reliable **bias and toxic content detection tools**.

With respect to data labelling, the lack of in-house expertise to create labelled datasets has increased the demand for **third-party data providers**. As such, the Global Data Collection and Labelling Market is accelerating at an impressive rate.[11] Online platforms such as Amazon's Mechanical Turk are also popular for **crowd-sourcing** campaigns for (trivial, non-expert) labelling tasks. These online platforms, however, are not useful when dealing with regional or lesser-spoken languages.

Businesses can hope to benefit from the forthcoming **European Data Governance Act**,[12] and the public sector has already begun to slowly benefit from the **Open Data Directive**[13] as evidenced by the European Language Resource Coordination (ELRC).[14] Similarly, As part of its European Digital Strategy, the European Commission recently published its **Data Act Proposal**,[15] which aims to "maximise the value of data in the economy by ensuring that a wider range of stakeholders gain control over their data and that more data is available for innovative use". The benefits of these policies can only be fully leveraged, however, if **sufficient awareness and engagement** levels at national level are reached. In fact, Berzins et al. (2019) report on the difficulties experienced across a number of EU member states in accessing public sector language data – due to the lack of awareness of the Open Data Directive. It may be the case, therefore, that solid national Open Data Policies that involve auditing procedures and Open Data Officers are also needed to help shift language data holders towards a data-sharing culture.

The creation of **Data Spaces** where companies can make data available for research under non-disclosure agreement terms (e. g., Smart Data Innovation Labs[16] SDIL or SDIL2) so far have not created a dynamic research ecosystem comparable with standard NLP data sets and models. However, under the EU Digital Europe Programme, new common *Data Spaces* are to be created that will "make accessible data across Europe, including information gathered from the re-use of public sector information, and become a data input source for AI solutions".[17] The spaces should be open to the public and private sectors. Such an approach to EU-wide data sharing is extremely promising in terms of addressing the current data accessibility gaps. To **ensure sufficient language coverage**, it is also hoped that Europe's lesser spoken and endangered languages will also be supported through incentives offered to governments, administrations, companies and citizens for donating language data.

5.2 Legal

Over the past several years, progress has been made in the research community with respect to cultivating a **culture of open data and data sharing**. Many top-tier publications require the release of datasets (where possible) in order to facilitate reproducibility of studies. Additionally most shared tasks (benchmark or evaluation campaigns) require a release of their specifically designed datasets for use by the wider research community (Escartín et al., 2021). These practices are only helpful however when related to datasets that are not **restricted by copyright, licensing agreements or privacy regulations**.

Since unconstrained, unstructured text can by its very nature often include personal data, **data protection and privacy** (DPP) policies can put limits on the type of data that can be made available for text analytics. **GDPR** (the EU's General Data Protection Regulation), while

[11] https://www.globenewswire.com/news-release/2021/11/01/2324173/0/en/Global-Data-Collection-and-Labeling-Market-Size-to-Grow-at-a-CAGR-of-27-7-from-2021-to-2030.html
[12] https://digital-strategy.ec.europa.eu/en/policies/data-governance-act
[13] https://digital-strategy.ec.europa.eu/en/policies/psi-open-data
[14] https://www.elrc-share.eu
[15] https://digital-strategy.ec.europa.eu/en/library/data-act-proposal-regulation-harmonised-rules-fair-access-and-use-data
[16] https://www.sdil.de/en/homepage
[17] https://www.eumonitor.eu/9353000/1/j4nvirkkkr58fyw_j9vvik7m1c3gyxp/vkp1fqrgymox

important for EU citizens' protection, significantly **hampers the extent to which language data can be sourced and reused** for machine learning based tools in Europe. The principles of DPP and legal provisions such as GDPR stipulate that data should only be used for a-priori defined narrow purposes and that these purposes must be made transparent to the data subject upfront. This proves problematic of course when dealing with induced models or datasets from web sources that have been reused without website owners' or individuals' consent. European-based researchers and LT developers cannot therefore use, share, modify or build upon many of these datasets – which sets DPP-compliant players in this field at a competitive disadvantage.

As the main issue related to GDPR restricted data concerns **Personal identifiable information (PII)**, steps have been taken recently towards developing tools that can **anonymise language data** in an attempt to overcome these barriers.[18] However, the task of anonymisation is difficult and does not always work with sufficient precision and reliability. Any text anonymisation in practice has to accept a potential residual risk of DPP non-compliance. Special usage rights have been called for to help advance NLP, particularly in domains where PII is prevalent in datasets (similar to the exemptions granted in the field of medical research).

5.3 NLU system limitations

In the ever-changing world of TA and NLU, some approaches are still in the early stage of adoption, while more advanced approaches still have much progress to make in terms of capabilities. Here we highlight some of the currently known gaps, draw-backs and areas for immediate improvement in this field.

Most of today's text analytics solutions are language-specific. **Various challenges** arise in many contexts (business, personal, governmental), where the **multilingual requirements** of customers and users from across Europe and around the globe need to be served. The adaptation of current technologies to a new language depends on a number of factors, not least the availability of training and evaluation data in that language when they are machine-learning based. For example, a conversational interface that needs to recognise intents expressed in 40 different languages, 40 different chatbots would need to be build, monitored and maintained. As we have seen, data availability is already a general problem, but when it comes to **lesser-spoken languages** with less digital content, this **scarcity is compounded**. If language agnostic tools are not a realistic goal, more innovation and investment are required in making this language adaptation process easier and less of a roadblock for LT providers, their customers, governments and the wider linguistically diverse public. This broadening of linguistic coverage can not solely rely on being market-driven (which is the main reason why relatively lesser spoken languages are being left behind).

Text Analytics is not only the process of analysing a source text sentence by sentence. Rather, key pieces of **contextual information** (i. e., pragmatics) such as the author, the intended audience, societal factors and the purpose of communication – the interactional and communicative context – need to also be considered. As such, there is much scope for improving contextualised and personalised analytics. One growing area of research is **multimodal NLP**, which aims to capture these contextual features and combine them with information elicted from text to make better judgements or predictions. For example, multimodal sentiment analysis which captures sentiment both through text and audio or visual data. As is the case for computational linguistics, such **interdisciplinary fields of research** require a broad amount of knowledge and expertise. As such, traditional silos of learning (e. g., third level institutions, training programmes) will need to **adapt and expand**.

[18] For example, the CEF-funded MAPA project set out to develop a toolkit for effective and reliable anonymisation of texts in the medical and legal fields for all official EU languages: https://mapa-project.eu

Data-driven approaches such as Machine Learning (and to a greater degree, Deep Learning) have been criticised for their **'blackbox'** nature. That is to say, when language data is converted to numeric or opaque vector representations in order to enable modelling or pattern inducing, it becomes difficult to (i) assess why a model is under-performing and (ii) overtly specify processing expectations of a system. Traditional symbolic AI (rule-based) approaches did not face these problems, but instead faced problems of scalability and robustness. Recent trends have emerged towards using **hybrid approaches** that can leverage the benefits of both **Deep Learning and Symbolic AI**. Likewise, innovative approaches have been developed to **injecting additional knowledge** into language models through techniques such as KnowBERT (Peters et al., 2019b) and K-Adapter (Wang et al., 2021a). These hybrid approaches have already proven to both increase performance in some settings and reduce the need for such large training datasets and as such deserve further exploration.[19]

One priority for many businesses and organisations is to build trust and confidence in these AI models. As a result, there has been a notable increase in attention given to the area of **Explainable AI**. In cases where decisions are made based on AI model prediction, it is important that businesses can assess these models' level of accuracy, fairness and transparency. One method of assessing what exactly a model is learning is the technique of 'probing' which has proved useful for improving some classification TA tasks such as parsing (e.g. Hall Maudslay et al. (2020)). There is therefore a clear need for a deeper understanding of these seemingly 'blackbox' NLP systems going forward, which will undoubtedly both increase users trust and guide further improvements.

Finally, further exploration is required into **extensibility methods** to include domain-specific knowledge (when large corpora are not available) and allow business or even LT providers to build custom extensions easily for machine-learning based systems.

5.4 Benchmarking

Benchmarking is the practice of establishing an evaluation reference point against which the performance of a system can be measured. Benchmarking campaigns, evaluation campaigns and shared tasks share the common objective of establishing standard datasets on which systems can be evaluated, establishing appropriate evaluation metrics and providing 'leaderboard' reports on best-performing systems so as to identify state-of-the-art (SOTA) performance.

In language technology (and NLU in particular), there is a **wide range of benchmarking frameworks** depending on the task at hand. **Evaluation metrics also vary** depending on the task, ranging from reporting on precision/recall and F1 scores for classification tasks, to exact matching/SacreBLEU[20] scores for dialogue systems. Current benchmarks in NLU include widely adopted ones like GLUE[21] and SuperGLUE.[22]

In academia, benchmarking is mainly used as a way to **advance research** (leaderboard-driven), while for industry it is a way to **determine the technical or market readiness** of a product. Moreover, savvy customers in this space will often set minimum accuracy scores in terms of the quality of the systems they require. While metrics and benchmarks exist for various TA sub-fields, it is often **difficult for users or buyers to determine how well their own content is or could be processed**. Often, custom code and data processing agreements need to be put into place before evaluating a solution (e. g., for entity linking on company-confidential data). Similarly, certain tasks are notoriously difficult to establish benchmarks for, such as information retrieval, as the relevance or non-relevance of a retrieved set of

[19] https://www.expert.ai/blog/symbolic-approach-nlp-models/
[20] https://huggingface.co/metrics/sacrebleu
[21] https://gluebenchmark.com
[22] https://super.gluebenchmark.com

documents can be wildly subjective, depending on the user.

In terms of the nature of datasets used in benchmarking, businesses require **realistic data** that is representative of the wide range of domains where TA and NLU systems are increasingly employed (e. g., health, manufacturing, finance, etc.). As such, the increasing trend for creating (often general purpose) synthetic data proves to be problematic. Some evaluation datasets is also often criticised in academic shared tasks, where they are sometimes referred to as "toy" examples that are not **applicable to real-world problems**. Therefore, in order to allow vendors, providers and suppliers to evaluate their own solutions, there is a clear **need for an increase in diversity, relevance and suitably annotated test data**.

Increased transparency is called for in general, both in terms of datasets used, metrics used and clarity provided on participation requirements (including within shared tasks (Escartín et al., 2021)). As with other areas of technology, **system certification** would hugely benefit the field of language technology.

In terms of enterprise data, another challenge pertains to the need for more comprehensive, standardized annotations and meta data[23] (so that, for example, data biases can be avoided more easily): as such, the **FAIR data principles** of making data 'machine actionable' should be applied (findability, accessibility, interoperability, and reusability).

Finally, the relevance of leaderboard positioning should be brought into question under certain circumstances. An important factor to consider is the trade-off between achieving the highest scores on a benchmarking leaderboard (and therefore setting a SOTA benchmark) and the **carbon footprint** of the energy consumption required to build these models (Strubell et al., 2019). Often, the little gain achieved through building larger models is relatively low with respect to the increase in harm caused to the environment. Those businesses achieving modest, yet adequate, performance quality can be overlooked in favour of outperforming systems that have significantly more computing power and a higher carbon footprint.

5.5 Investment protection and interoperability

There has been a significant move towards **open-source tooling** for language technologies, particularly with the emergence of repositories such as Github[24] and online AI communities such as Hugging Face[25] which provide platforms for easy of sharing in the LT research and development sphere. As a result, many NLU system components are available for a 'plug-and-play' interaction with complex pipelines during software development. In terms of academic research or open-source development, interoperability is mostly facilitated, especially if common libraries are used or datasets are in formats such as JSON. However, in the absence of standards, **interoperability at an enterprise level** can prove to be more challenging when proprietary software or data formats are part of the mix.

Language technology often requires significant investments by users and buyers. Three sample investment areas are language assets, annotated test suites, and services meshes – solutions that include amongst others, technical services provided by language technologies. A language asset like an enterprise scale terminology database, ontology, or authoring memory can easily require costly months or years of work. The same holds for the other two investment areas. Accordingly, TA solutions need to be built with investment protection and interoperability in mind. Otherwise, **risks such as vendor lock-in are likely to surface**.

Additionally, **official standards** (e. g., from the World Wide Web consortium (W3C) like for provenance[26]), or industry-standards (e. g., from schema.org for vocabularies related

[23] For example schema.org, DCAT, PROV-O, VOID
[24] https://github.com
[25] https://huggingface.co
[26] https://www.w3.org/TR/prov-o/

to entity types[27]) are important ingredients for protecting investments since they facilitate interoperability and reuse. Standards often also assist during solution development since they embody knowledge from experts, and are very often accompanied by an ecosystem of tools/libraries (see e. g., SPARQL[28] for operations related to semantic representations). Thus, existing standards should guide for example any cataloguing, data annotation or Application Programming Interfaces (APIs) (e. g., the use of BCP47-based language tags, W3C Internationalization Tag Set, and support for task-specific formats such as OASIS XLIFF (for bilingual data), or ONNX for neural networks).[29]

5.6 Conformance

A special dimension related to standards concerns conformance. "Conformance is the fulfillment of specified requirements by a product, process, or service."[30] While such requirements are not so crucial for academic research, they are highly relevant to enterprise language technology development as they **assure quality standards for consumers**. Accordingly, requirements statements are needed for any TA artefact. For entity detection, this requirement statement could for example mention that a conformant application must be able to detect any of the entity types of the Common Locale Data Repository[31] in Spanish and Portuguese[32] In particular, in the context of regulated industries, certification – the assignment of a label based on transparent testing, and compliance with conformance criteria – may need to be considered.

5.7 Consumer-grade tools for domain experts

Today, most work in the ML-driven LT ecosystem **requires expert level skills** in the realm of tools related to data management, data science and NLP processing. This **creates bottlenecks** since it does not allow domain experts (e. g., experts in finance) to become actively involved without rather extensive tool training, and without the need for understanding the underlying technology. The 'design' of this ecosystem also causes overhead and delays since work between tool experts (e. g., data scientists) and domain experts needs to be coordinated. For example, a considerable length of time may pass between a domain expert selecting suitable data for a use case, and the first evaluation of this data in an NLP-related process.

What is lacking as a way to address this is the availablity of consumer-grade, highly usable, possibly **low code/no code tools for domain experts, or even citizen scientists**. Ideally, these datasets and NLP tools should be developed in collaboration with usability research. Such tooling would allow domain experts to play a more active role in the development of solutions for application scenarios they are familiar with. They would be able to:

- Generate, label, access and process structured data/knowledge sources (e. g., knowledge graphs) without needing to be experts in the underlying technology (e. g., SPARQL queries).
- Generate, use and evaluate TA systems (e. g., using recall and precision to evaluate the accuracy of term linking solutions).

[27] https://schema.org
[28] https://www.w3.org/TR/sparql11-overview/
[29] https://onnx.ai
[30] https://www.w3.org/TR/qaframe-spec/#specifying-conformance
[31] https://unicode-org.github.io/cldr-staging/charts/latest/by_type/index.html
[32] https://www.w3.org/TR/its20/#conformance and http://docs.oasis-open.org/xliff/xliff-core/v2.1/os/xliff-core-v2.1-os.html#Conformance for sample conformance clauses.

6 Text Analytics: Contribution to Digital Language Equality and Impact on Society

Today, text analytic tools can help societies and individuals in various ways by supporting tasks that involve the discovery of information (facts, rules, and relationships) in text. As we have seen there are widely-used and indispensable applications available to businesses, consumers, citizens and governments that cover a wide range of usage scenarios, starting from recommendation and sentiment analysis tools to intelligent virtual assistants, business intelligence tools, predictive analytics, fraud management, risk management, and cybercrime prevention. Text analytic tools are also widely used in online and social media data analysis, which is of use to both businesses and governments.

Currently, however, all of these advances and digital innovations are really only supporting those who speak major and well-resourced languages (e. g., English, French, German, Spanish). Adapting these technologies to support other languages across Europe is not a trivial task of simply localising software or connecting existing technology to local databases or information sources. Languages differ significantly in many ways, not just in words but also inflectional nature (e. g., plural forms of nouns or tenses of verb), sentence structure (word order), idiomatic uses, semantic variability, and so on. To that end, these applications need to be built upon systems that understand the underlying patterns in each language that requires support. As today's AI powered NLP techniques are data-driven, this means that sufficient amounts of data need to be made available in order to adapt technologies to these languages. However, it is not always a case of plugging in new data-sets to existing technologies. Due to the fact that languages and domains can differ so significantly, various parameter tuning, system adaptation or hybrid implementation may also be required to achieve robust and reliable technologies.

As it stands, the availability and quality of text analytic (TA) tools differ from language to language, from task to task and domain to domain. This is mainly as a result to the different levels of investment that have been made into TA technologies across various languages. Market demand (number of speakers) often drives investment by technology companies, which explains why languages such as English are so well supported. Additionally, governments of some economically advanced countries have invested well into R&D efforts to help support their official language, e. g., Spanish. Where this investment is lacking or non-existent for a language, it is rendered unsupported and lacking in terms of speech and language technology. As a result we currently find a striking imbalance with respect to digital language support across Europe.

It is possible to address this imbalance through TA and NLU, in a step towards achieving digital equality for economies, societies and language communities. Multilingual Europe and globalisation requires tools that can process and analyse texts published in languages other than just the major languages. Therefore, to reach the goal of digital equality text analytic tools need to understand and provide access to text, regardless of the language in which it is written.

6.1 Text Analytics Tools for Digital Language Equality and Multilingual Europe

TA and NLU can play a major role in overcoming current language and technology barriers that prevent the flow and accessibility of information and knowledge across Europe. Simply from an economic perspective, this language barrier has an impact on the Digital Single Market (European Parliament, 2018). Europe's Single Market seeks to guarantee the free movement of goods, capital, services, and people. The role of technology in this is key as countries seek to ensure continued access to this single market and information such as

product information, national and local policies, education information, trade information, financial information, and so on. Such information needs to be accessible to all EU citizens. Text analytic tools (together with machine translation solutions and other cross- and multilingual solutions) are the key elements for accessing this information and knowledge across Europe.

The 2012 META-NET White Paper Series (Rehm and Uszkoreit, 2012) reported on an analysis of language technologies and resources available for EU languages. The results showed that with respect to TA, *Good support* only applied English and *Moderate support* to five widely spoken languages - Dutch, French, German, Italian and Spanish, leaving other 24 (out of 30) European languages of this study in a cluster of *Fragmented, Weak or no support*.

Today, all 24 EU official languages benefit from basic underlying TA tools – tokenizers, lemmatizers, morphological analysers, part-of-speech tagging tools, and syntactic parsers (for details see also Table 3, Section 4). While the quality of reliability or robustness of these tools varies across languages, their existence represents a step in the right direction. On the other hand, more sophisticated TA tools and services, e. g., summarisation tools, are available only for a small number of languages (Table 4, for details see also Table 1 and Table 2, Section 3, as well as Table 3, Section 4).

Language	Chatbots and conv. syst. build.		Entity Extraction		Sentiment Analysis		Summarization	
	Global	ELE	Global	ELE	Global	ELE	Global	ELE
Bulgarian	0	2	1	16	0	9	0	2
Croatian	0	2	2	16	0	6	0	1
Czech	1	4	1	13	0	7	0	2
Danish	0	3	2	18	1	12	0	4
Dutch	2	14	3	34	4	22	0	7
English	3	25	6	78	6	54	2	19
Estonian	0	2	1	11	0	6	0	3
Finnish	0	4	2	21	0	14	0	5
French	3	13	6	42	5	24	2	10
German	3	42	6	59	5	40	2	13
Greek	0	2	2	17	0	10	0	5
Hungarian	0	3	2	16	0	9	0	3
Irish	0	0	0	3	0	1	0	0
Italian	3	6	6	28	5	22	2	8
Latvian	0	3	1	10	0	5	0	0
Lithuanian	0	3	0	7	0	5	0	0
Maltese	0	0	0	5	0	1	0	0
Polish	0	7	2	17	0	17	0	5
Portuguese	2	4	5	25	5	20	1	9
Romanian	0	3	2	16	0	8	0	4
Slovak	0	4	3	13	0	7	0	2
Slovene	0	3	1	14	0	7	0	2
Spanish	3	7	6	51	5	43	2	14
Swedish	0	4	3	22	1	15	0	5

Table 4: Language support of text analytics services for 24 EU languages: by global technology providers and reported to ELE survey.

It can be clearly seen that there is a large number of languages already at a disadvantage

due to little or no existence of such tools (e. g., Lithuanian, Irish, Maltese and Slovene). Low numbers can also indicate early stages of research and development for a given language in these areas. Therefore, the quality or reliability of many of the tools accounted for should also be considered when trying to gauge a clear picture of TA support. In terms of Irish for example, a small sentiment lexicon exists but there is no sentiment-tagged corpus or reliable sentiment analysis tool available.

The figures presented in Table 4 are indicative of the broad disparities that exist across languages in terms of the types of TA support currently provided. "Support" describes whether such a tool exists that can handle, understand and process text written in a given language. As these common tools are crucial for many applications used by businesses, governments and citizens, it becomes clear how societies and economies can be negatively impacted by their lacking.

Some of the main reasons why most of sophisticated TA techniques are not available for many EU languages (Rehm et al., 2020) are **lack of data and data sparsity (especially for morphologically rich languages)** for training and testing TA technologies and the **complexity of technology adaptation and transfer in low resource settings**. For instance, in the case of dialog systems and chatbots, analysis of available datasets for dialog modeling clearly demonstrates gap of language resources for less resourced languages (Serban et al., 2018; Leonova, 2020).

Recent techniques used to extend TA tools to less resourced languages include the use of large pre-trained language models and zero-shot transfer. For instance, with help of multilingual BERT models (Devlin et al., 2019), solutions for POS tagging, named entity recognition and dependency parsing, could potentially be extended to less resourced languages that only have small training sets. The DLE problem and its implementation into truly multilingual internet is also addressed by Horizon 2020 project EMBEDDIA[33] by leveraging innovations in the use of cross-lingual embeddings and deep neural networks to allow usage of monolingual resources across languages, in particular, less and low resourced languages.

6.2 Impact on Society

With the democratization of artificial intelligence and the development of accurate and clever AI solutions that communicate with users in natural language, AI technologies already affect business activities, society and individual users' lives. From an economic perspective, Gartner (November, 2021) forecasts the worldwide artificial intelligence (AI) software revenue to total $62.5 billion in 2022, an increase of 21.3% from 2021.[34] With respect to areas of high impact, the top five use-case categories for AI software spending, according to Gartner, in 2022 will be knowledge management, virtual assistants, autonomous vehicles, digital workplace and crowdsourced data.

Intelligent, AI-based, virtual assistants are already in demand in the digital market and use of them in the workplace is growing. Gartner (August, 2020) predicts that by 2025, 50% of knowledge workers will use a virtual assistant on a daily basis, up from 2% in 2019. For public sector and businesses this means an opportunity to use an intelligent virtual assistant technology to take care of more repetitive and auxiliary business processes. By 2030, Gartner predicts that the decision support/augmentation will be the largest type of AI accounting for 44% of business value, while agents representing 24%.[35] These predictions of course only hold for countries with lesser-spoken languages if the technology is there to support them.

[33] http://embeddia.eu

[34] https://www.gartner.com/en/newsroom/press-releases/2021-11-22-gartner-forecasts-worldwide-artificial-intelligence-software-market-to-reach-62-billion-in-2022

[35] https://www.gartner.com/en/newsroom/press-releases/2019-08-05-gartner-says-ai-augmentation-will-create-2point9-trillion-of-business-value-in-2021

If not, it is clear how an economic divide will emerge, as countries with sufficient language technologies will gain advantage.

The following are examples of how TA technologies can have a positive impact in our world today as governments, businesses and consumers. Also highlighted are the negative impacts that the lack of such technology will eventually have societies and economies that may be left behind digitally.

6.2.1 Governmental and Public Services

Today, some Government organizations already apply NLP solutions to help them deliver efficient public services and improve governance. According to the Gartner Digital Transformation Divergence Across Government Sectors survey[36] chatbots are leading the way in government NLP and AI technology adoption - 26% of government respondents reported that they have already deployed them, while 59% are planning to have deployed them within the next three years. In the case of machine learning-supported data mining – only 16% have currently deployed it with a further 69% planning to do so within the next three years.

In the case of governmental organisations, one of the key challenges faced is obtaining relevant information from huge volumes of unstructured text. In these cases, TA tools can be used to: help to solve routine tasks (e. g., with help of virtual assistants many common citizen information-related questions could be answered without human intervention), improve public services (e. g., through analysis of public feedback or engagement), assist process analysis (e. g., identifying potential risks, investigating or enhancing policy analysis) or even address critical government issues. Adding to this, is the extra layer of complexity faced when dealing with data containing PII or protected under GDPR that cannot always be shared with researchers or developers due to its sensitive nature.

In the past several years, European governments have begun to access EU funding through international projects that support improvements in the public sector by harnessing the power of AI-powered TA and NLU. These projects are often a consortium of government agencies, academic institutions and industry partners. One such example is the Connecting Europe Framework project NLTP,[37] which sees the creation of a novel state-of-the-art, AI-driven National Language Technology Platforms for several European countries – Croatia, Estonia, Iceland, Latvia and Malta, weak or no support in 2012, according to the META-NET White Paper Series (Rehm and Uszkoreit, 2012). NLTPs will provide national public administrations and SMEs with mature, tightly integrated LT services to enable multilingual access to information and online services by combining the most advanced language technology (LT) tools and solutions developed in the CEF-AT and other European and national programmes. In case of Malta the English language has become the default language of choice across most technological devices, thus the NLTP will support the curation of Maltese language tools and resources and ensure the presence of the currently low-resourced Maltese language in digital environments (Cortis et al., 2021). In Latvia for example, *Hugo.lv* is a Latvian state administration language technology platform that is freely accessible to every resident of Latvia. Today *Hugo.lv* provides automatic translation, speech recognition and speech synthesis, as well as a range of tools for supporting multilingual features in e-services. *Hugo.lv* is customized to the Latvian language and state administration documents, thus, its quality is much higher than in other online NLP services.

Another noteworthy CEF-funded project is MAPA (Multilingual Anonymisation for Public Administrations).[38] The MAPA project is taking steps towards addressing such PII obstacles

[36] https://www.gartner.com/en/newsroom/press-releases/2021-10-05-gartner-says-government-organizations-are-increasing-

[37] National Language Technology Platform, 2020 CEF Telecom Call – Automated Translation (CEF-AT)

[38] https://mapa-project.eu

facing LT in the public sector by leading the development of a toolkit for effective and reliable anonymisation of texts in the medical and legal fields in all EU official languages.

NLP also can help governments engage with citizens and provide answers to their questions. For instance, Hugo.lv includes the catalogue of virtual assistants (VA) developed by public bodies. Currently there are more than 10 VAs for different public services, including the Bank of Latvia, the State Revenue Service, the Register of Enterprises and many others. In order to make it simpler for users to consume information, particular attention is paid to information visualization using infographics, tables, images and videos, etc.

The text analytic solutions are critical for defense and intelligence sectors as they help to improve predictions. For example, RED (Real-time Early Detection) Alert project,[39] analyzed social media conversations to counter terrorism and to provide early alerts of potential propaganda and signs of warfare, while DARPA applied NLP tools to improve efficiency of defense analysts (Eggers et al., 2019).

6.2.2 National Interests

At a national level, governments are employing AI powered technology to ensure that their nations continue to evolve at an equal pace in today's digital world. In many ways, the processing and understanding of text is fundamental to this progression at national level. From the perspective of national security and integrity, TA and NLU is often applied to flag or identify possible risks that can be detected in written format. National concerns such as threats to national security, money-laundering and people-trafficking are often intercepted through advanced technology in this space. When relevant documents or texts are written in technologically unsupported languages however, such instances of national interest remain undetected.

Similarly, new advances have been made in the area of event detection, based on what is being reported in real time in social media by citizens and eye-witnesses (e. g., natural disasters, accidents). Of course, this analysis on large amounts of data is only possible for the content in languages that are supported sufficiently through TA. Where a language is not supported, any relevant content written in that language is therefore disregarded and rendered unusable.

Court and criminal justice systems are now benefiting from multimodal approaches to content retrieval combining speech processing and NLU to assist in the discovery of evidence amongst large amounts of unstructured audio and video content. Inequalities are likely to arise in the legal system however, as processing times will improve only for those whose languages are suitably supported through these technologies.

Sentiment analysis of online political commentary (eg. news articles, social media, etc.) is often used by governments and political parties to gauge their popularity based on the electorate's opinions online (i. e., what is being said about them). In addition and true to predictions[40] that the future of government service ratings would lie in the hands of sentiment mining, the UK is one such example of a government who has embraced the power of topic modelling and sentiment analysis to analyse the feedback provided by citizens in their GOV.UK website.[41] Similarly, online data mining is used as a technique for predicting election outcomes. However, in a multilingual society, only the opinions or comments of those in the technologically supported languages will be represented. In other words, the voices of many will be left unheard, unrepresented and unaccounted for.

In terms of national media, a political bias classifier has been recently developed for German that can assist in identifying left or right-leaning content (Aksenov et al., 2021). In times

[39] https://redalertproject.eu
[40] https://datasmart.ash.harvard.edu/news/article/from-comment-cards-to-sentiment-mining-301
[41] https://dataingovernment.blog.gov.uk/2016/11/09/understanding-more-from-user-feedback/

of growing polarisation, tools such as this can help to ensure more balanced reporting which in turn can prevent potential divide across communities.

6.2.3 Education

School based learning and education is changing rapidly in terms of technological support. In many learning environments, there is a shift away from the traditional pencil and copybook approach towards technology supported learning. This shift is supporting learning and growth, and ultimately improving quality of lives and leading to better societies.

For instance, Computer Assisted Language Learning (CALL) tools are increasingly employing TA and NLU to create intelligent learning support systems. For example, personalised or adaptive learning is a technology that allows the identification of a student's progress and gaps in their knowledge, while adapting the curriculum, learning pace or learning goals to suit the learner. Such adaptive learning has proven invaluable for subjects such as language learning, maths and science (Chen et al., 2021).

Bilingual countries often feature a more dominant language that influences the language medium through which education is offered across society. In these cases, language immersion schools are also offered to those who, instead, want their children to receive an education through their mother tongue. While these lesser-spoken language-medium schools are key to ensuring continued use of the language across generations, the availability or lack of language technology to support learning could eventually create a divide in the levels of education on offer to citizens, contributing further to inequalities.

At secondary and tertiary level, students no longer consult libraries and encyclopedias for supplementary knowledge, but instead employ information retrieval to access additional online educational resources or content such as Wikipedia. The accessibility of these resources rely on the availability of both proofing tools and information retrieval tools that can match keyword searches to relevant content. Without them, students will be disadvantaged in their learning capabilities.

For those with writing difficulties or mobility issues, there are now tools to make learning easier and more accessible. Some of these are multimodal involving automatic speech recognition tools (ASR) in conjunction with transcription tools that remove the need for these users to type. Without these technologies, a clear divide will appear with respect to how those disabilities will be supported across language communities.

Virtual learning assistants and augmented virtual reality (e. g., for language learning immersion), like adaptive learning, rely on TA and NLU technology to guide and support students. This type of learning if often effective when the teacher-student ratio is low and students require additional support.

A major challenge for assessing large groups of students is the ability to track learning progress among them. Learning progress analytics is being made possible through TA and NLU, in settings such as automatic scoring as applied to English content in the US.[42] Very little research has progressed to market for these types of applications for other languages.

6.2.4 Career and Growth Opportunities

The world of job-seeking and career moves has changed significantly over the past several years. Today, in the English speaking world at least, professional networks and job databases such as LinkedIn have changed the way in which recruiters find potential candidates and job seekers find potential career options. TA and NLU are fundamental in this process and much of the language technology powering these kind of systems is AI-driven. In many ways, they also benefit from the power of knowledge graphs and relationship linking to enable the right

[42] https://www.ets.org

recruiters find the right candidates by matching users' CVs to job descriptions. This provides an advantage to both businesses and individuals.

Upskilling and re-education is also high in demand nowadays, with learning platforms providing tailored learning based on users' interests, previous experiences and so on. These personalised systems are also enabled through TA technologies, matching the right courses with the right users. Such learning platforms are therefore enabling growth and opportunity that will improve not only the lives of individuals but also leading to wider impact at a society level as a result of a strengthened and more skilled workforce.

In the absence of varied language support in this sphere, it is evident that only specific language communities (businesses and citizens alike) are set to gain advantage through a more skilled workforce.

6.2.5 Digital Interaction and Connected Societies

Language support and proofing tools (e. g., spell-checker, grammar-checker, auto-correct, predictive text) facilitate more efficient and seamless creation of digital text content. Today, it is unusual to find (for English at least) a platform or app that does not provide such language support (e. g., customer review forms, micro-blogging platforms such as Twitter, blogs, messenger tools, etc.). As such, they are often viewed as a fundamental requirements for any text-based content creation technology. However, such support does not always extend to other languages. Consider the case where a user tries to write content in their own language, yet there is a lack of such support for that language. Their words are instead "auto-corrected" to a word in another supported language or underlined in red as a typo or invalid word. This is a frequent occurrence and challenge for speakers of minority languages. In such cases, one of two outcomes occur: (1) over time, users will default to writing in another supported language (if they can speak one) or (2) they will stop using the technology. In the case of (1), this is a clear step towards language shift and eventual language decline, particularly amongst younger generations. In the case of (2), this creates a divide in levels of accessibility and usability across language communities.

Similarly, the availability of proofing tools also influences a society or community's connectedness. While speech technology is becoming more prevalent in Business to Business (B2B) and Business to Customer (B2C) interactions, much of our personal interactions with each other still rely on language technologies that facilitate written communication (e. g., emails, online social networks, instant messengers, chat rooms, etc). As this continues to be the trend, we can see clearly how, through the lack of basic technological support, a language community could not continue forging or strengthening these connections through their own language. Such scenarios inevitably leads to disconnect and possible divide.

6.2.6 Health

One major challenge health systems face worldwide is the large amounts of data that has been collected relating to patients, and the inability to parse or processes this data efficiently. When the data is in text format, TA and NLU can be used to create links between patients diagnoses, patient records, recent medical research, and so on. recognising medical terms and named entities is crucial for this, particularly if knowledge graphs are used as a solution.

The global companies — Google, IBM, Microsoft, and Amazon — today provide healthcare-focused text analysis services. For example, Text Analytics for health[43] by Microsoft is one of the features offered by Azure Cognitive Service for Language, allowing to extract and label relevant medical information from unstructured texts. The service can be used for many

[43] https://docs.microsoft.com/en-us/azure/cognitive-services/language-service/text-analytics-for-health/overview?tabs=ner

different types of unstructured medical documents, such as discharge summaries, clinical notes, clinical trial protocols, medical publications and more. The TA service performs NER, relation extraction and entity linking.

TA tools support clinical decisions by providing easy and efficient access to health related information. With help of TA tools personnel can review massive quantities of unstructured clinical and patient data and identify candidates for clinical trials. TA tools also help in clinical documentation process by creation of electronic health records (EHR) from audio records and extraction of necessary information.

According to the Health Europe,[44] virtual cognitive assistants could drastically reduce the administrative burden and lead to improved patient experience and health outcomes. Already in the medical industry we can see investment in cognitive agents like virtual medical billing assistants, virtual radiology assistants, virtual plan of care assistants, virtual medical testing assistants, etc.[45]

According to ResearchAndMarkets,[46] the virtual medical assistant market is expected to grow from $1.1 billion in 2021 to $6.0 billion by 2026. The smart speakers segment of the healthcare virtual assistants market should grow from $813.1 million in 2021 to $4.4 billion by 2026, while chatbot segment - from $317.3 million in 2021 to $1.6 billion by 2025.

During the height of the Covid-19 pandemic, the role of virtual assistants increased in medical domain, since VAs were able to provide the public with convenient and fast access to trustworthy information. Using natural language understanding techniques, e. g., intent detection and named entity recognition, these chatbots can retrieve or can generate answers to common questions, such as the latest regional, national and international illness statistics, relevant contact information including information hotline numbers, information about the virus, border crossing, the nearest analysis delivery points, how to act in various situations etc. Several Covid-bots were been developed that collected and presented accurate and validated information from different national and international sources, including the World Health Organization (WHO).[47] Another group of Covid-bots provides means for COVID-19 auto-diagnosis.[48]

Multilingual and cross-lingual text analytic tools for medical domain can also help in knowledge transfer, fact finding and fast solution finding when rare and less common information is necessary. This is particularly relevant if solution needs to be provided in urgent situations, where immediate response is crucial. For example, text analytics proved helpful during the COVID-19 pandemic in exploring massive amounts of international data on the virus and making it accessible to medical professions worldwide.

A growing area of research and development in the health domain is the emergence of medical transcription tools that will support doctor-patient interactions. Research has shown that these interactions lack in terms of the attention the doctor can spend engaging with the patient face-to-face, due to the overhead of note-taking. Medical transcription or scribe tools, using a combination of speech and NLU technologies, are being introduced to improve this interaction and also make note-taking more consistent and structured. The quality of the data then captured through these tools will further lead to improvements in healthcare.

[44] https://www.healtheuropa.eu/patient-experience-virtual-cognitive-assistants/91679/

[45] A Review of Cognitive Assistants for Healthcare is recently published by Preum et al. (2021)

[46] ResearchAndMarkets.com

[47] WHO's Health Alert interactive service on Facebook Messenger is available in English, French, Spanish, and Arabic; Estonian and English Suve bot – https://eebot.ee; German – https://covidbot.d-64.org; Latvian – https://covidbots.lv

[48] For example, Finnish, Swedish and English – https://koronabotti.hus.fi; French – https://maladiecoronavirus.fr; Italian – https://covidbot.d-64.org

6.2.7 Business and Consumer Benefits

All European economies have seen a shift towards eCommerce in the past several years. This shift has benefited both businesses (wider market reach) and consumers (convenience, more choice). TA plays an important role in supporting both parties. From a commercial perspective, businesses no longer need to conduct polls to gauge customer satisfaction, Instead they can use sentiment analysis to assess online reviews, mentions in social media and customer feedback forms. Personalised advertisement also helps to find the right potential customer base.

From a customer's perspective, more efficient customer service (through chatbots, virtual assistants or automatically generated FAQ sections) makes buyer-seller interactions more seamless. Multilingual systems widens these benefits even further. Effective online search through product websites is also supported through TA.

From an EU Digital Single Market perspective, the importance of being able to reach wider markets and consumer bases should not be underestimated. Nor should the importance of effective multilingual online dispute resolution.

It is clear therefore, that for economies and societies to grow and evolve at the same pace, they need the same level of access to such TA tools.

7 Text Analytics: Main Breakthroughs Needed

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 and lately with very large pre-trained language models. Indeed, NLP practitioners find themselves today in the midst of a significant paradigm shift in NLP. 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, 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 NLP. Neural language models and related techniques are key to NLP progress and therefore being able to build them for target languages with the same quality as English is key for language equality.

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 leading language resource repositories, platforms, libraries, models and benchmarks have begun to make inroads in this regard. 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 NLP for less developed languages.

7.1 Sufficient resources

In recent years, the NLP community is contributing to the emergence of powerful deep learning tools and techniques that are revolutionizing the approach to NLP tasks. We are moving from a methodology in which a pipeline of multiple modules was the typical way to implement NLP solutions, to architectures based on complex neural networks trained with vast

amounts of data. This rapid progress in NLP has been possible because of the confluence of 4 different research trends: 1) mature DL technology, 2) large amounts of data (and for NLP processing large and diverse multilingual textual data), 3) increase in HPC power in the form of Graphic Processing Units (GPUs), and 4) application of simple but effective self-learning and transfer learning approaches using Transformers (Devlin et al., 2019; Liu et al., 2020b; Torfi et al., 2020; Wolf et al., 2020). Thanks to these recent advancements, the NLP community is currently engaged in a paradigm shift with the production and exploitation of large, pre-trained transformer-based language models (Han et al., 2021; Min et al., 2021a).

As a result, various IT corporations have started deploying large pre-trained neural language models in production. For instance, Google and Microsoft have integrated them in their search engines. 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 language understanding tasks. However, despite their impressive capabilities, large pre-trained language models raise severe concerns. Currently we have no clear understanding of how they work, when they fail, and what emergent properties they present. Some authors call these models "foundation models" to underscore their critically central yet incomplete character (Bommasani et al., 2021). There are also worrying shortcomings in the text corpora used to train these anglo-centric models, ranging from a lack of representation of low-resource languages, to a predominance of harmful stereotypes, and to the inclusion of personal information. Moreover, these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud computing time, and environmentally, due to the carbon footprint required to fuel modern servers with multiple GPU hardware. This also means that only a limited number of organisations with abundant resources in terms of funding, computing capabilities, NLP experts and corpora 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 research groups can afford modern NLP research (Ahmed and Wahed, 2020). Thus, this transformative technology poses important concerns from a research, innovation but also environmental perspective. To tackle these questions, much critical interdisciplinary collaboration and research are needed.

In summary, there is a lack of necessary resources (experts, data, computing facilities, etc.) compared to large US and Chinese IT corporations (Google, OpenAI, Facebook, Baidu, etc.) that lead the development of these new NLP systems. In particular, the "computing divide" between large firms and non-elite universities increases concerns around bias and fairness within this technology breakthrough, and presents an obstacle towards democratizing NLP. In fact, in the EU there is an uneven distribution of resources (funding, open data, language resources, scientists, experts, computing facilities, IT companies, etc.) by country, region and language. Thus, the development of these new NLP systems would not be possible without sufficient resources, as well as the creation of carefully designed and constructed evaluation benchmarks and annotated datasets for every language and domain of application. Finally, we note with concern a tendency to focus on state-of-the-art results exclusively with the help of leaderboards, without encouraging deeper understanding of the mechanisms by which they are achieved. We believe that this short-term goals can generate misleading conclusions and, more importantly, can direct resources away from important efforts that facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical and unbiased language understanding and communication. Progress in these fields will help creating transparent digital language equality in Europe in all aspects of society, from government to businesses to the citizens.

7.2 Natural Language Understanding

Current language models contain billions of parameters and are pre-trained using thousands of millions of multilingual documents. As such, pre-trained language models are shown to encode a large amount of background knowledge, which allow them to obtain meaningful representations of the text or generate documents from a given topic. 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.

Recent work has shown that pre-trained language models can robustly perform NLP 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; Ding et al., 2021). *Prompting* is a technique that involves adding a piece of text (called *promts*) to the input examples to encourage a language model to bring to the surface the implicit knowledge you are interested in, hence helping the language model to perform the task at hand. Surprisingly, fine-tuning pre-trained language models on a collection of tasks described via instructions (or prompts) substantially boosts zero-shot performance on unseen tasks (Wei et al., 2021; Sanh et al., 2021; Min et al., 2021c; Ye et al., 2021; Aghajanyan et al., 2021; Aribandi et al., 2021; Tafjord and Clark, 2021; Lourie et al., 2021). Thus, this is a very active area of research, in which scientists try to use different configurations and prompts for both augmenting the input examples, or verbalizing the desired outcomes of the language model. Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pre-trained language models, prompt learning and self-supervised systems opens up the way to leverage NLP techniques for less developed languages.

Integrating commonsense in NLP systems has long been seen as a near impossible goal —until recently. Now, research interest has sharply increased with the emergence of new benchmarks and language models (Mostafazadeh et al., 2016; Talmor et al., 2019; Sakaguchi et al., 2020; Ma et al., 2021; Lourie et al., 2021). This renewed interest in common sense is encouraged by both the great empirical strengths and limitations of large-scale pretrained neural language models. On one hand, pretrained models have led to remarkable progress across the board, often surpassing human performance on leaderboards. On the other hand, pre- trained language models continue to make surprisingly silly and nonsensical mistakes.[49] This motivates new, relatively under-explored research avenues in commonsense knowledge and reasoning.

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. It is worth investigating possible opportunities to leverage both structured and unstructured information sources and to enhance contextual representations with structured, human-curated knowledge Peters et al. (2019b); Colon-Hernandez et al. (2021); Lu et al. (2021).

However, despite claims of human parity in many of the NLP tasks, Natural Language Understanding (NLU) is still an *open research problem* far from being solved since all current approaches have *severe* limitations. Language is grounded in our physical world, as well as in 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

[49] https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

ingredients to achieve human-level NLU (Storks et al., 2019). Thus, 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). Following the irruption of deep learning methods, new paradigms have been adopted and the field of NLU has advanced significantly in the last few years.

7.3 NLP systems and humans working together

While NLP systems based on deep learning obtain remarkable results on many tasks, the output provided by NLP models, particularly those models that generate text, are still far from perfect. For example, the textual snippets generated by advanced language models such as GPT and successors is formed by syntactically correct sentences that seem to talk about a particular topic. However, when analyzing the sentences, it is clear that there is a lack of semantic coherence among them. Humans are still needed to monitor the output of automatic NLP systems and, possibly, adapt them to the task at hand.

Traditional linear NLP development pipeline is not designed to take advantage of human feedback. Advancing on the conventional workflow, there is a growing research body of Human-in-the-loop (HITL) NLP frameworks, or sometimes called mixed-initiative NLP, where model developers continuously integrate human feedback into different steps of the model deployment workflow. This continuous feed- back loop cultivates a human-AI partnership that not only enhances model accuracy and robustness, but also builds users' trust in NLP systems (Wang et al., 2021c).

This form of human intervention when developing NLP tasks is not new, and has been used for a long time in, e. g., Machine Translation, where automatically generated translations are often post-edited by humans. In areas such as text simplification or summarization, machines have shown a reasonable capacity to recognize the most salient points of large documents, but have problems turning these points into coherent texts. But having a machine highlight the key ideas in a document and a human turn that into a short snippet, outperforms either working alone.

In the foreseeable future we expect this interaction to be higher, as AI and NLP become embedded in everyday work processes.

7.4 Open-source culture will strengthen the NLP field

The NLP field is overall committed to the open-source culture, and this commitment is expected to continue in the future. On the other hand, the aspect of replicability, that is, the ability to replicate the results reported on a scientific paper, is a very central topic in NLP. Nowadays the majority of scientific papers are often accompanied with the source code (often distributed on platforms like github) and data that is required to replicate the experiments. Leaderboards such as NLP-progress,[50] Allen Institute of AI leaderboard[51], Papers with code[52], or Kaggle[53] are meant to encourage participation and facilitate evaluation across many different NLP tasks and datasets.

Recent progress in NLP 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[54] may open up these advances to a wider

[50] http://nlpprogress.com
[51] https://leaderboard.allenai.org
[52] https://paperswithcode.com/area/natural-language-processing
[53] https://www.kaggle.com/datasets?tags=13204-NLP
[54] https://huggingface.co

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).

As a result of this commitment, the NLP community has considerably increased the access to models and datasets, which are nowadays publicly available and easily accessible. This culture focused towards sharing fosters opportunities for the community to inspect the work of others, iterate, advance upon, and broaden access to the technology, which will in turn strengthen the collective skill sets and knowledge.

The culture is also reflected in the industry, where companies like MonkeyLearn aim to democratize NLP and machine learning technology, allowing non-technical users to perform NLP tasks that were once only accessible to data scientists and developers. MonkeyLearn's point-and-click model builder makes it easy to build, train, and integrate text classification or sentiment analysis models in just a few steps, which means we can expect to see more and more businesses implementing NLP tools in 2021.

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.[55] 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.

7.5 Broadening the NLP field

As the technology involving NLP becomes more mature, valuable synergies will be created among related AI and Machine Learning disciplines. Deep learning techniques and, particularly, transfer learning approaches, have considerably flattened the learning curve that ML and AI practitioners suffer when approaching problems that deal with text. As a consequence, ML and AI practitioners with relatively few background knowledge in language technologies will be able to contribute to the field.

The arrival of practitioners from related fields will broaden the NLP field, which have been traditionally formed by highly niche specialists. This will in turn often many opportunities to develop systems and applications that integrate information from many modalities, such as text images, speech or video. For example, Samsung is combining NLP with video imagery to help self-driving cars interpret street signs in foreign countries. NLP and computer vision are also combined to develop applications that help translate video to text for accessibility purposes, improve descriptions of medical imagery, etc.

In general, we can say that this cross-pollination of fields will be highly beneficial for NLP and will allow creating fruitful synergies among different disciplines.

Regarding the negative aspects, there is a risk of underestimating the specific characteristics and idiosyncrasies needed to develop NLP systems. Text analysis requires a deep understanding of underlying linguistic processes that textual elements undergo, which is a complex process full of small details and nuisances, and that when underestimated can cause NLP applications to fail.

[55] https://bigscience.huggingface.co

8 Text Analytics: Main Technology Visions and Development Goals

Nowadays artificial intelligence is part of our lives. We use it when browsing the internet, using our smartphone, shopping on the internet, or interacting with smart devices and appliances. In the future, artificial intelligence will permeate new applications and domains, making people relations with machines more human-like. So, human-machine interaction needs to be more personalised, fluid and nuanced. Language processing is a key technology to reach this level of communication between humans and machines.

There is no doubt that in the last decade natural language processing and understanding has experimented a great leap forward thanks to advances in deep learning and supporting hardware, which has allowed training models from very large text collections, something that was not possible before. This new paradigm in natural language processing has opened new possibilities for text analysis and understanding but also presents some drawbacks that need to be addressed in order to ensure wider adoption of the technology. Also, deep learning systems need to coexist with knowledge-based systems, also referred as symbolic systems, for natural language processing that have existed and been used in real-life applications for many years. One of the main development challenges for natural language processing is to get the best from deep learning and symbolic systems while minimizing their respective drawbacks.

In this section, we provide an overview of the main technology visions for natural language processing and understanding between now and 2030. We have identified developments for increasing the language support of such technologies, putting people needs in the center of any breakthroughs involving language technologies, the integration with other modalities of information in addition to text, the hybridization process for symbolic AI and neural systems, and the need of a new generation of benchmarking tools. We finalize presenting some future application scenarios that illustrate what shall be possible in 2030 once these developments are available.

8.1 Multilingual Text Analytics

Language support beyond widely spoken languages, including minority and under-resourced languages, is still a pending issue in text analytics and natural language understanding. The investment of language technology providers in such languages seems inhibited most likely due to a comparatively lower profitability compared to mainstream languages, considering the number of potential users of the technology.

Nevertheless, the current trend in language technologies relying on neural language models and research on unsupervised, and zero-shot learning opens new possibilities to increase the coverage of minority and under-resourced languages in the text analytics industry. Language models have shown promising results in zero-shot setting in a wide range of tasks (Radford et al., 2019; Brown et al., 2020; Gao et al., 2021). That is, language models learn to perform task from patterns naturally occurring in text, eliminating or reducing to a great extent the need of additional labeled data which is a scarce resource for many languages.

We expect that the language coverage of text analytics tools will be enhanced thanks to a mixture of research breakthroughs on multilingual language models (Conneau and Lample, 2019), language agnostic models (Aghajanyan et al., 2019), and neural machine translation (Johnson et al., 2017). Research on these subjects is underway and show promising results, even for under-resourced languages (Conneau et al., 2020), paving the way for truly multilingual language technologies.

8.2 Human-centric Language Technologies

In the last years generalist language models have excelled in language processing tasks leveraging the vast amount of linguistic knowledge learned from general or domain-specific text collections. Nevertheless, so far neural language models are mainly addressed as a one-size-fits-all approach, offering almost no customization beyond the user-generated data used to fine-tune (Devlin et al., 2019) or prompt (Brown et al., 2020) the models for downstream tasks. The current research lines focused on unsupervised and zero-shot learning (Gao et al., 2021) in natural language processing delve into this issue since users of the technology have little to say in the learning process.

Moreover, the dominant data driven approach to natural language processing and the race for accuracy has yielded opaque tools that are hard to interpret and biased tools that perpetuate social stereotypes on gender and racial basis found in text collections (Sheng et al., 2019). The lack of transparency makes it difficult to build trust between users and system predictions, having negative consequences in technology adoption. Biased tools have a direct impact in society as a whole and can have a negative impact on marginalized populations (Sheng et al., 2021).

We advocate for a next generation of language processing tools that care about end users' needs and expectations, making them part of the design and learning process. These tools will be human-aware, emotional, and trustworthy, avoiding bias, offering explanations, and respecting user privacy. Moreover, human intelligence will be used on pair with machine learning techniques to produce better language technologies. Human feedback can serve as a guide on the learning process telling the machine what users want and what they do not want (Christiano et al., 2017). Reinforcement learning from human feedback is a promising research avenue (Stiennon et al., 2020; Li et al., 2016) to use human intelligence to improve language processing tools. Also, interactivity with domain experts and users, as per Shapira et al. (2021) and Hirsch et al. (2021), is a key area for further advances beyond the usual supervised paradigm.

8.3 Neurosymbolic/Composite AI/Hybrid language technologies

In the NLP community, there is a certain tension between machine learning-based methods and those that advocate for a structured or symbolic knowledge-based approach. Some believe that the statistical approach is too data hungry and does not lead to real understanding of the meaning of text. Others think that the symbolic approach is too rigid and that it requires a substantial effort to write ontologies and rule-based systems that cover all corner cases. In the end, both positions have some truth, but the real problem is that, by focusing on the limitations of one approach or the other, we run the risk of missing the unique opportunities that each of them has to offer.

As practitioners come to realize the inevitable limitations of purely end-to-end deep learning approaches, which increase in the case of underrepresented languages (both in terms of available pre-trained language models and suitable training corpora for such languages), the transition to hybrid approaches involving different ways to combine neural and symbolic approaches to NLP becomes an alternative that appears more and more tangible. Therefore, it is important that we exhaustively discuss the components necessary to build such systems, how they need to interplay, and how we should evaluate the resulting systems using appropriate benchmarks.

The field of neurosymbolic approaches to NLP and NLU, recently rekindled in the context of the Semantic Web Hitzler et al. (2019) among other areas, will be increasingly important in order to ensure the integration of existing knowledge bases within our models, as already shown by approaches like KnowBert Peters et al. (2019a) and K-Adapter Wang et al. (2021b). Not only to make NLU models aware of the entities contained in a knowledge base and the

relations between them from a general purpose point of view, as provided by resources like Wikipedia, but when it comes to quickly incorporating pre-existing resources from vertical domains and custom organizations into our models in a cheap and scalable way.

Some argue Sheth et al. (2017); Shoham (2015); Domingos (2012) that knowledge graphs can enhance both expressivity and reasoning power in machine learning architectures. Others Gómez-Pérez et al. (2020) propose a working methodology[56] for solving NLP problems that naturally integrate symbolic approaches based on structured knowledge with neural approaches. These are the first practical steps in this direction. Many more are needed, particularly in a multilingual and language equality scenario. In doing so, it is particularly important to show not only that integrating structured resources and neural approaches to NLP and NLU is possible, but that it is also practical and desirable for successfully solving many real-life problems, using specific examples and success stories.

8.4 Multimodal AI

Text analytics tools are specialized on extracting useful information and insights out of written language. However, such specialization is typically unaware of other modalities of information often used along text, such as images, audio, and video, which can enrich the analytics process. For example, advertisement comes with images of the products being sold while product reviews include the review descriptions and pictures or videos uploaded by consumers.

Different modalities can be combined to provide complementary information that may be redundant but convey information more effectively (Palanque and Paternò, 2000). Thus, analytic tools that can analyze jointly different modalities and extract information from them can carry out a more comprehensive analysis. Furthermore, multimodal analysis has allowed machines for the first time ever to pass a test from middle school science curricula involving questions where it was necessary for the model to understand both language and diagrams in order to answer such questions (Gomez-Perez and Ortega, 2020).

This convergence across modalities requires synergies from AI research fields that until now are being conducted individually such as natural language processing, automatic speech recognition and computer vision. Deep learning techniques will play an important role in multimodal analysis. Neural networks have been successfully used to process text, speech and images, and architectures initially proposed for one modality of information have then been adapted to other modalities. Recently, transformer architectures (Devlin et al., 2019), initially proposed for natural language processing, are being used for image processing (Dosovitskiy et al., 2021) and for cross-modal information processing including images and text (Hu and Singh, 2021). Similarly, in the past, convolutional neural networks (CNNs) followed the inverse path, from computer vision (Krizhevsky et al., 2012) to text processing (Tay et al., 2021).

Other approaches based on contrastive language–image pre-training, like CLIP (Radford et al., 2021), emphasis the present and future relevance of the zero and few-shot scenarios. CLIP shows that scaling a simple pre-training task is sufficient to achieve competitive zero-shot performance on a great variety of image classification datasets by leveraging information from text. The approach uses an abundantly available source of supervision based on pairs of text and images found across the internet, resulting in a gigantic language-vision dataset. Unfortunately, the text in all such pairs is in English only, showing how language inequality also impacts on language-vision tasks. Investment in multilingual resources that extend datasets like CLIP will also be necessary to make this type of technology available across all the European languages as well as underrepresented languages in general.

[56] Methods, resources and technology on Hybrid NLP https://github.com/expertailab/HybridNLP

8.5 Benchmarking

Benchmarking aligns research with development, engineering with marketing, and competitors across the industry in pursuit of a clear objective. However, evaluation for many natural language understanding (NLU) tasks is currently unreliable and biased, with plenty of systems scoring so highly on standard benchmarks that little room is left for researchers who develop better systems to demonstrate their improvements. The recent trend to abandon independent and identically distributed (IID) benchmarks in favor of adversarially-constructed, out-of-distribution test sets ensures that current models will perform poorly, but ultimately only obscures the abilities that we want our benchmarks to measure. Adversarial data collection, understood as the process whereby a human workforce interacts with a model in real time, attempting to produce examples that elicit incorrect predictions, does not meaningfully address the causes of model failures, as shown, e. g., by Kaushik et al. (2021) for question answering.

Restoring a healthy evaluation ecosystem will require significant progress in the design of benchmark datasets, the reliability with which they are annotated, their size, and the ways they handle social bias. Even more so when we expand our view to a multilingual landscape, such as the European multilingual reality. Recent work lays out different criteria that future NLU benchmarks should meet and argue that most current benchmarks in NLU including widely adopted ones like GLUE[57] and SuperGLUE[58] fail at addressing such criteria. Along these lines, Bowman and Dahl (2021) propose four criteria that we would like our benchmarks to satisfy in order to facilitate further progress towards the vision of building machines that can demonstrate a comprehensive and reliable understanding of everyday natural language text, including language variety, and vertical domains. Among language understanding tasks, special emphasis is placed on those that use labeled data and that are designed to test relatively general language understanding skills, for which the design of benchmarks can be especially difficult.

Such criteria, which are particularly suitable for the digital language equality scenario, cover the following dimensions:

- **Data validity and specificity.** An evaluation dataset should i) reflect the full range of linguistic variation in the relevant domain, context, and language variety, ii) plausibly test all the language related behaviors we can expect for the task, and iii) be sufficiently free from annotation artefacts. The method used to collect the data (naturally-occurring examples, expert authored examples, crowdsourcing, adversarial) is utterly important. One promising direction involves methods that start from relatively high-quality crowdsourced datasets and then use domain experts to augment the datasets in order to mitigate annotation artifacts.

- **Reliable annotation is critical.** There are three main annotation failure cases that should be avoided: i) examples that are carelessly mislabeled, ii) examples that have no clear correct label due to unclear task guidelines, and iii) examples where annotators systematically disagree. Possible ways to deal with these issues include treating ambiguously labeled examples in the same way as mislabeled examples, systematically identifying and discarding them during a subsequent validation phase.

- **Statistical significance, complexity and cost.** Benchmark evaluation datasets should be large and challenging enough, but the costs of doing so can be prohibitive. For example, for the task of natural language inference, labeling an existing sample requires a minimum 45” of work by a crowdworker, while creating one example from scratch takes at least 1’. Assuming an average of 15 €/hr pay rate, a ten-way dataset of 500.000

[57] https://gluebenchmark.com
[58] https://super.gluebenchmark.com

examples would cost over 1M €, or more if more experience or careful annotators are used. Coming up with ways to scalably and sustainably produce such datasets in an analogous way to modern manufacturing chains is a challenge that needs to be addressed. Approaches like gamification to motivate and involve annotators could provide free human labor capacity, but it comes at the cost of defining the annotation task as a game that is attractive to play.

- **Disincentives for Biased Models.** A benchmark should, in general, favor a model without socially-relevant biases over an otherwise equivalent model with such biases. Many current benchmarks fail this test. Because benchmarks are often built around naturally-occurring or crowdsourced text, it is often the case that a system can improve its performance by adopting heuristics that reproduce potentially-harmful biases, as shown by Rudinger et al. (2017). Developing adequate methods to minimize this effect will be challenging. While it would be appealing to try to guarantee that evaluation data does not itself demonstrate evidence of bias, there is currently no sign of robust strategies for reliably accomplishing this goal. As illustrated by Gonen and Goldberg (2019), work on the closely-related problem of model bias mitigation has been fraught with false starts and overly optimistic claims. More promising alternatives shall involve the use of additional, expert-constructed test datasets and metrics, each of them isolating and measuring a specific type of bias.

Furthermore, as advocated by several panels of experts including Church et al. (2021), much more emphasis will need to be given to typical realistic settings, in which large training data for the target task is not available, like few-shot and transfer learning. Moreover, while measuring performance on held-out data is a useful indicator, held-out datasets are often not comprehensive, and contain the same biases as the training data, as ilustrated by Rajpurkar et al. (2018) inter alia. Recht et al. (2019) also showed that this can lead to overestimating real-world performance. Approaches like CHECKLIST, proposed by Ribeiro et al. (2020), advocate for a methodology that breaks down potential capability failures into specific behaviors, introducing different test types, such as prediction invariance in the presence of certain perturbations and performance on a set of sanity checks inspired in software engineering.

Finally, benchmark design shall fit realistic data compositions, rather than synthetic ones within the comfort zone. Addressing such shortage of benchmarks for real-life scenarios will require the decided involvement of industry, including both providers of language technologies, customers, adopters, and practitioners, to collaboratively develop with academia reference benchmarks for key NLU tasks. Two requirements must be compulsory for such benchmarks: On the one hand, they will need to cover a representative sample of the key sectors in the European economy, including among others finance, health, tourism, manufacturing, and the corresponding added value chains. On the other hand, such benchmarks need to be born multilingual and cover each of such economic sectors for each of the European languages, guaranteeing language equality regardless of the size of the market associated with each language.

8.6 Future Technology Scenarios

In this section we present future usage scenarios enabled by next generation text analytics tools once the main technology vision and development goals described above have become a reality.

8.6.1 Virtual Multilingual, Multimodal Scientific Agent

In 2030 researchers are challenged by the huge number of scientific publications that keeps growing fast paced every day. The scientific production of countries with tradition on the scientific enterprise has been increased with the contributions from big emergent economies where research and innovation has become a priority. This publication deluge makes it very hard for researchers to keep track of major breakthroughs in their field. In addition, the use of English as the International Language of Science keeps being a burden for countries where English is not a native language, particularly for those with a low rate of English speakers. To help researchers in their scientific endeavors a new Virtual Scientific Agent VSA has been released incorporating the latest advances in language processing including multimodal abstract summarization, emotional conversational AI, Multimodal QA (Knowledge base, Table/List, Texts, Visual), and Composite AI.

A young researcher commissioned to investigate on marine litter detection in the Spanish Mediterranean coast could use the VSA as follows:

- Agent, I need to start new research on Marine Waste detection
- VSA: Very interesting problem, pollution in the oceans is an ecological disaster
- VSA: Marine litter is mostly made of plastics, are you interested on micro-plastics, macro plastics, or in marine litter in general?
- I didn't know that, what percentage of the waste is plastic?
- VSA: 80% according to Wikidata, and this peer-reviewed paper found in OpenAIRE[59]
- Thanks. I need a brief report of the state of the art in marine plastic detection.
- VSA: Your report is ready; I have included most influential papers and papers with highest positive impact in social media in the last 5 years. I also included diagrams showing the detection processes and arranged the publications in a Field of Study taxonomy. In the summary you'll find the state-of-the-art technique, its performance metrics, and improvements over the previous state-of-the-art.
- Thanks, now I would like to review latest news about pollution in the Spanish Mediterranean coasts ...

At the beginning of this dialogue the VSA uses emotion AI to recognize that a new research activity implies a lot of effort, and the researcher could feel overwhelmed by the task, hence it motivates her by emphasizing the importance of the problem. Then the VSA provides more information about marine litter to increase the user confidence in its knowledge about the subject. To do so, the VSA uses its generative language model to produce pairs of questions and answers related to marine litter and chooses one question, e. g., "what is the main source of marine plastic?", to use the answer as additional information. This new and unknown data to the researcher, makes her want to know more about the percentage of plastic in marine litter. To answer this question the VSA uses its composite QA model relying on knowledge graphs (e. g., Wikidata) and large scientific publication repositories (e. g., OpenAIRE).

Next, the researcher asks for a report on the state of the art on the subject. The VSA uses its multimodal summarization module to generate the report combining paper descriptions, images, and information from tables in leaderboards. Finally, the multilingual capabilities of the VSA are leveraged to search for local news about pollution in the Spanish coast and generate a summary in English for the researcher.

[59] https://explore.openaire.eu

8.6.2 Personalised, Multimodal Educational Content for Societal Advancement

Digitisation has revolutionised content production, monetarisation, dissemination, consumption and more. Yet, digitization for at least one kind of content has still a long way to go: educational material. Educational material currently is published mostly on paper. In many countries, there is a small but focused industry with a long tradition on creating books and related material conforming to country and region-specific requirements of educational content.

The covid-19 pandemic served as a wake-up call for this industry: It changed learning settings from "physical" to "virtual", or "hybrid". It amplified the demand for "real" digital educational material – material that goes far beyond text and includes multimodal content (some if offered via dedicated interactive apps). Most learners– pupils, graduate and undergraduate students, employees – use the Internet as their main channel for education.

At the same time – partially originating from global competition – the need and aspiration for (lifelong) learning has grown substantially. New information that is also relevant for education (e. g., about political developments, progress in science, changes in society) is shared rapidly on the Web but takes years until it is covered in schoolbooks in the languages needed. The rapid growth of information, and the increasing velocity of new information needs challenges learners since it becomes harder to find adequate material: in their language and learning context – for example considering external aspects like country, learning grade, curricula etc., as well as personal aspects like learning history.

Text analytics, text and data mining, natural language understanding, and other language technology fields (most notably machine translation, and speech technologies) appear to be prerequisites to address the challenges related to learning content, no matter what (mix of) modalities are involved. They can provide, or contribute to functionalities such as the following:

- Transform existing learning content into (at least partially) language agnostic representations.
- Generate learning content from language agnostic representations of knowledge.
- Translate learning content into the language of the learner.
- Adapt content "on demand" to a learner's language proficiency.
- Encode the language and learning context of students, so that relevant educational content can be found.
- Encode a concrete learning situation, including the emotional state of a student, to choose, e. g., larger vs. smaller and easier vs. more challenging learning units (cf. Adaptive Learning Environments)
- Create virtual teachers (especially for remote communities)
- Create multimodal immersion learning.

The realisation of these opportunities requires a collaboration of diverse constituencies:

a. Researchers/commercial technology providers in the realm of language technologies (and artificial intelligence in general) to contribute state-of-the-art technology and to advance it even further.

b. The established industry of educational publishing, which provides large volumes of educational content and expertise about didactics / (state) regulations for accreditation of educational content.

c. Providers of IT solutions that manage the whole life cycle for digital educational products, including ERP systems (e. g., for production planning, and finance), content management systems, delivery systems, learning management systems and more. A crucial role will be played by providers of IT solutions that support new business models: models that take into account the shift from selling books series with a development and production cycle of several years, to creating and selling content items instantly, in the language of the users(s) and based on their learning context and learning situation.

8.6.3 Multilingual Human-like Interaction for Inclusive, Human-centric Natural Language Understanding

Today text analytic tools help societies and individuals process and analyze huge volumes of information mostly in widely spoken languages and in many cases in a monolingual way. Therefore there is a urgency for text analytic and natural language understanding techniques for less resourced European languages. By 2030 text analytic tools need to support human-like interaction to access the overall European knowledge regardless of language. To reach this goal text analytic tools and methods need enable natural language understanding in low resourced settings at similar level as for resource rich languages.

The multi-/cross- lingual text analysis tools that operate over overall European knowledge could be used in any domain that require text analysis, knowledge extraction and natural language understanding. For example, in healthcare, it will allow to analyze patent data, case studies, etc. in any language and summarize analysis results in user's language. In case of education, it will provide access and suggest the most appropriate education materials at pan-European level. For customer service, it will allow to analyze data from different countries, or, in case of multilingual societies input in any language.

Expected impact includes:

- This technology will foster inclusiveness of smaller communities by enabling human-like interaction with ICT solutions for speakers of smaller European languages in their native language.
- European text analysis tools for NLU equally supporting all European languages will help reducing technology gap between well-resourced and less resourced languages.

9 Text Analytics: Towards Deep Natural Language Understanding

Much has been said in recent times about the expected impact of intelligent systems in many aspects of our lives. Today's large amount of available data, produced at an increasing pace and in heterogeneous formats and modalities, has stimulated the development of means that extend human cognitive and decision-making capabilities, alleviating such burden and assisting our drivers, doctors, teachers and scientists, and sometimes even replacing them. In scientific disciplines like biomedical sciences, some like Kitano (2016) even propose a new grand challenge for this kind of systems: to develop an artificial intelligence that can make major scientific discoveries and that is eventually worthy of a Nobel Prize. Though still far from realization, this scenario suggests the time is ripe for a shared partnership with machines, whereby humans can benefit from augmented reasoning and information management capabilities if machines are endowed with the necessary intelligence to assist with such tasks. Through such partnership, we can expect a virtuous circle of training data collection, active learning, and interactive feedback, which will result in self-adaptive, ever-learning systems.

We have already seen signs of such partnership, for example in the application of generative language models like GPT-3 to produce text given a prompt, with applications in a wide variety of business sectors. Based on these developments, some suggest[60] that the future of artificial intelligence lies in the development of systems that allow maintaining a conversation with a computer. This scenario goes beyond current chatbot technologies, which many deemed as mere digital parrots, able to copy form without understanding meaning, but nevertheless capable of creating a dialogue with the user. This is something that often seems missing from the introduction of AI systems like facial recognition algorithms, which are imposed upon us, or self-driving cars, where the public becomes the test subject in a potentially dangerous experiment. With AI writing tools, there will be the possibility for a conversation. However, this will require advances in knowledge representation, true understanding of meaning and pragmatics, and the ability for the models to explain and interpret their predictions in ways that humans can understand and relate to.

The artificial intelligence community and particularly the areas related to text understanding will soon need to address other issues like fairness in ways that tangibly and directly benefit disadvantaged populations. We have spent large amounts of effort discussing about fairness and transparency in our algorithms. At the algorithmic level, fairness has to do with the absence of bias in the models that, e. g., in natural language understanding are used to address tasks that may range from the evaluation of mortgage applications or insurance policies to medical examination and career recommendation. If the algorithms are biased, then so will be the outcome of their predictions and inequalities would be perpetuated as the use of artificial intelligence unrolls more and more deeply in society.

This is essential work, but now it is time to develop systems and tools that have a tangible impact in business and society. The lack of resources in a specific language to train a natural language understanding model in such language is another source of discrimination. A very visual example in a related domain has to do with the use of a smartphone navigation app in a wheelchair – only to encounter a stairway along the route. Even the best navigation app poses major challenges and risks if users cannot customize the route to avoid insurmountable obstacles. Similarly, the lack of availability of service functionalities in all language will have an undesired effect in the respective populations. Accessibility, education, homelessness, human trafficking, misinformation, and health among others are all areas where artificial intelligence and text understanding can have a major positive impact on people's quality of life. So far, we have only started to scratch the surface.

10 Summary and Conclusions

Text analytics and natural language understanding (NLU) deal with extracting meaningful information and insights from text documents of any kind as well as to enable machines to understand such content in depth, similar to how a human would read a document. Corresponding tools and applications have been on the market for several years and have successfully found applications in many sectors including Health, Education, Legal, Security, Defense, Insurance, and Finance to name but a few. However, existing text analytics and NLU services do not cover all languages equally as of today.

The market offer around these technologies tends to gather around languages that cover a larger segment of the population, maximizing the return of investment. As a consequence, there is a risk of discrimination in terms of the coverage provided to European languages with a lower number of speakers despite current efforts to ameliorate this situation. To reduce the coverage gap across languages both in the market and in society, technical, regulatory, and societal advances are required that increase access to text analytics and NLU

[60] https://www.theverge.com/22734662/ai-language-artificial-intelligence-future-models-gpt-3-limitations-bias

technologies regardless of the specific European language and territory.

In this document, we present a comprehensive overview of text analytics and NLU under the perspective of digital language equality in Europe. We focus both on the research that is currently being undertaken in foundational methods and techniques, as well as gaps that need to be addressed in order to offer improved text analytics and NLU services in the market across languages.

Before going in detail with all the conclusions of our analysis, we emphasize two points that in our opinion will be particularly critical to ensure digital language equality in Europe:

- Neural language models and related techniques are key to sustain progress in Language Technologies. **Therefore, being able to build neural language models for target languages with the same quality as English is key for language equality**.
- Multilingual data is the key element to train such models in the target languages. We should not take for granted that large amounts of publicly available corpora of good quality can be readily obtained for all European languages, rather the contrary. **The effort to ensure that all languages have large amounts of publicly available corpora of good quality, taking into account fairness issues, should be at the center of any future efforts for digital language equality.**

Next, we summarize the main takeaways and provide a synthetic list of eight guidelines and recommendations related to text analytics and NLU in the context of European digital language equality.

1. **Language equality in text analytics is a transformative and integrative force for social good.** We have shown examples of how language equality in text analytics and NLU can stimulate development in such important aspects for our societies as access to health, public administration services for everyone, better education, and more business opportunities. These will contribute to more developed societies, which in turn can contribute to progress and prosperity, creating new markets for text analytics and other areas related to artificial intelligence and language technologies Europe-wide. The Latvian case perfectly illustrates this situation. However, this is not a widely spread situation across all European languages, yet. The question we should make ourselves is what is the alternative? what will the social cost be if the required policies do not effectively reach all European languages between now and 2030? Since text analytics and NLU are pervasive to our lives these days, those societies that fail in taking up this challenge and lack access to such services in their own languages will also risk being in serious disadvantage for future development. Language, as the main human communication mechanism, is key in this regard.
2. **The balance between legitimate commercial interests and equal access to opportunities is fragile** when it comes to digital language equality in text analytics. Our analysis shows how global providers of text analytics services tend to concentrate their offering (and associated investment) in widespread languages, neglecting the long tail of languages that have a comparatively smaller population of speakers. Category A of European languages is reasonably well covered. However, for languages in category B, global players offer scarce support to no support at all. Interestingly, the situation is better with languages in category C, probably because they originate from other territories outside of Europe that happen to be of strategic interest for these players. In contrast, European initiatives like ELG provide a much more equitable coverage. From this analysis two reflections emerge. First, it is a European business to ensure that all European languages are properly covered. Therefore, European companies in the text analytics space should benefit from incentives that allow them to focus on such

languages. Such incentives should naturally come from a thriving market demanding these services in Europe first of all, but also in other forms that could translate, e. g., into tax breaks for both European technology providers and European customers that acquire their services to address less represented languages. Second, this effort should involve European technology providers but, in order to create traction, also consumers of such services at the different levels of the European public administration and large European companies should be active participants as well.

3. **Possible incentives to language equality in text analytics and NLU are not only financial.** Actually, the research communities in natural language processing and computational linguistics are in the middle of a heated debate around these topics. Research publications in these areas should not take for granted that contributions evaluated only in the English language are equally valid for all languages. In this regard, the so-called Bender rule originally proposed in Bender (2011) calls researchers to "Do state the name of the language that is being studied, even if it's English". Acknowledging that we are working on a particular language foregrounds the possibility that the techniques may in fact be language specific. Conversely, neglecting to state that the particular data used were in, say, English, gives [a] false veneer of language-independence to the work. (Bender, 2011, p. 18)." This should result into the generalized practice of naming the languages studied, the practice of asking, as a reviewer, which languages were studied, and the practice of being skeptical of claims of language independence when only one test language was used. Such practice should be of particular relevance for European research venues.

4. **Neural language models are a corner stone of most NLU and text analytics pipelines now and in the next years.** However, current methods to generate such language models are hardware-intensive, require large amounts of text data to train them, and such training comes at the cost of high energy consumption and a large carbon footprint. Because of this, most of the neural language models available nowadays, like BERT, RoBERTa, T5, GPT-3, etc., have been trained on general purpose documents collected from the internet and freely available resources, which hinders their application in vertical domains, requiring additional pre-training on relevant data that is not easy to find.

   Few are the examples of neural language models pre-trained on domain-specific data. Plus, those available so far like Space RoBERTa, recently released by Berquand et al. (2021) in the domain of Space science and engineering, tend to be in English to maximize adoption, without either multilingual or language-specific versions. An alternative would be to invest in domain-specific, vertical and multilingual collections of text data that covers the strategic sectors of European economy and administration for training truly domain-aware, multilingual neural language models.

   Another complementary approach consists of looking at pre-existing structured resources including multilingual knowledge bases and knowledge graphs, some of them multilingual, and inject the knowledge contained in such resources into pre-trained language models. We have discussed this hybrid approach and some of the steps that are already being taken in this direction as one of the technology visions highlighted in this document.

   Finally, another way to ameliorate this situation can be to encourage European research efforts to look into methods that allow the creation of native-born multilingual language models. Such strategy should be supplemented with continued research on approaches that enable the portability of existing pre-trained language model representations from largely represented languages to underrepresented languages without requiring to re-train from scratch.

5. **Data is key.** Without data, NLU models and text analytics solutions based on machine learning approaches cannot be trained. However, suitable data and particularly multilingual text is hard to find and expensive to annotate in order to enable subsequent fine tuning of pre-trained language models on downstream tasks like classification, sentiment analysis, entailment, question answering, etc., which lie at the core of many text analytics services. While much progress has been done in creating large-scale labeled data sets for majoritarian languages, it is not feasible, especially from a business-driven perspective, to do this for the literally thousands of languages spoken on the planet, including all European Languages. In this report, we summarize some of the most successful techniques in the last five years relevant for language equality, multilingual language models and transfer learning between languages.

   Self-supervision enables computers to learn about the world just by observing it and then figuring out the structure of images, speech or text. Having machines that do not need to be explicitly taught to classify documents and images or understand language is simply much more scalable. Therefore, further research and development work on self-supervised and semi-supervised training approaches will be required. Furthermore, few-shot and zero-shot scenarios where hardly any task-specific training data or no data at all is available for fine tuning will be increasingly common as we address the needs of under resourced languages. One of the most promising directions according to the scientific publications in the last two years in Language Technologies is few-shot learning using prompts. However, all this research has been done for English alone. Key contributions and surveys on these topics include Liu et al. (2021) and Min et al. (2021b). Techniques that enable cross-lingual approaches to leverage models trained in better represented languages will also be increasingly important.

   Nevertheless, learning from data using supervision or self-supervision ignores the human knowledge about the task at hand and the user needs and expectations are constrained to the data used to train the models and the model's learning objective. As a result, when model predictions are misaligned with the user expectations the only way to influence the predictions is with more data. Reinforcement learning offers alternatives to train models on learning objectives more closely related to the end user needs. Applied to text analytics and NLU, reinforcement learning is an interesting research line to align users' expectations and model predictions. In addition, human knowledge encoded in structured knowledge graphs can be used to enhance learning models, yielding more robust text processing tools that leverage the strengths of deductive and inductive reasoning.

   As suggested in the previous point on neural language models, there is little or no doubt that enough general-purpose data can be collected in the different European languages that will suffice to pre-train language models for each of such languages following self-supervised approaches. The problem comes in satisfying the needs of domain and task specific data to adapt such models, building upon them text analytics solutions that solve real-life problems in each of the different business sectors and languages. We can visualize this as a matrix, where the rows are European languages, the columns are the European business and societal sectors of interest, and the cells are the specific datasets available. Accomplishing digital language equality and realizing its economic potential depends to a large extent on the success of populating the cells of such matrix.

6. **However, data tends to be locked in regulatory and corporate silos.** Research and solutions for language technologies that address problems of business and social relevance is underdeveloped. A major reason is lacking enterprise data available to researchers in industry and academia. As enterprise data is by nature confidential and companies need to respect data protection regulations, the barriers for making data

available are high. The idea to create data spaces where companies can make data available for research under NDA terms still need to crystallize into a dynamic research ecosystem that can be compared to generally available text analytics and NLU datasets and models.

To address this bottleneck further collaboration is required between industry, academia and European institutions that facilitates the creation of multilingual text data spaces across the different strategic business sectors. Part of this work would benefit from an optimal balance between European regulations like the GDPR and the use of data for research purposes. Currently, companies that abide to GDPR have a significant disadvantage to those that do not. In the end, to be competitive, European companies may need to use large neural language models built by third parties in USA or China that do not necessarily adhere to the same principles.

As proposed earlier in the document, a possible remedy could be special data usage rights. Text analytics and NLU research and development could be exempted or have lower barriers in relation to GDPR regulations. We are aware that this should be debated between the different European stakeholders involved. However, this type of measures would be aligned with existing GDPR exceptions for research in other areas like medicine. The European data spaces initiative recently launched by the European Commission as part of the Digital Europe programme offers an exciting perspective in addressing some of these challenges, as recently announced.[61]

7. **Benchmarking is broken and needs to be fixed and updated.** Evaluation for many NLU tasks is currently unreliable and biased, with plenty of systems scoring so highly on standard benchmarks that little room is left for better systems to demonstrate their improvements. The recent trend to abandon traditional, independent and identically distributed benchmarks in favor of adversarially-constructed, out-of-distribution test sets ensures that current models will perform poorly, but ultimately only obscures the abilities that we want our benchmarks to measure.

   Restoring a healthy evaluation ecosystem, particularly one involving a vision for digital language equality, will require significant progress in the design of benchmark datasets, the reliability with which they are annotated, their size, and the ways they handle social bias. However, it is key if we want to make well-grounded progress that supports not only technical but also ethical an societal issues As suggested above, benchmark design shall fit realistic data compositions, rather than synthetic ones within the comfort zone. Addressing such shortage of benchmarks for real-life scenarios will require the decided collaboration of European industry and academia.

8. **Text does not live in isolation. Information is cross-modal.** As shown in this document and elsewhere, text is rarely found in isolation in real-life. This is both challenging but also fortunate: text understanding can be used to provide a better analysis of images, video and audio, as well as the other way around. Many tasks in artificial intelligence, such as image and video captioning and machine reading comprehension, are actually cross-modal and represent an important and developing area of research and innovation. Addressing many of the market and societal challenges towards digital language equality illustrated in this document will benefit from taking into account the cross-modal scenario to leverage additional sources of free supervision. In this regard, recent advances like OpenAI's CLIP and Meta's Data2Vec (Baevski et al., 2022) seem promising. Such approaches propose solutions based on zero-shot and self-supervised learning that have exciting possibilities when it comes to deal with the scarcity of (particularly, annotated) data, demonstrating the relevance of such free supervision coming

[61] P. Gelin. A Language Data Space for Europe. META-FORUM 2021 https://youtu.be/RhpdB34zJcs

from other modalities in addition to text, like audio, images or video. Interestingly, all such models are currently available in English only.

To conclude, within artificial intelligence the field of text analytics and natural language understanding has an enormous potential to impact the development of business and society. At the same time, addressing the challenges related to linguistic discrimination and language barriers to communication and free flow of information is an utmost priority for Europe. Our hope is that this document, summarized in the guidelines and recommendation distilled in this final section, provides the necessary insight, from the point of view of text analytics and natural language understanding technologies, to the definition of the strategic roadmap between now and 2030 towards full digital language equality within the European Union. In doing so, we aim at contributing to establishing a fair, inclusive and sustainable Multilingual Digital Single Market based on equality, where text analytics and NLU have a decisive role to play, acting as a multiplier of opportunities and collaboration among key European stakeholders, including academia, industry, public administration, and 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.

Armen Aghajanyan, Xia Song, and Saurabh Tiwary. Towards language agnostic universal representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4033–4041, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1395. URL https://aclanthology.org/P19-1395.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.468.

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

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.

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.

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julián Moreno-Schneider, and Georg Rehm. Fine-grained Classification of Political Bias in German News: A Data Set and Initial Experiments. In Aida Mostafazadeh Davani, Douwe Kiela, Mathias Lambert, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem, editors, *Proceedings of the 5th Workshop on Online Abuse*

and Harms (WOAH 2021), pages 121–131, Bangkok, Thailand, 8 2021. Association for Computational Linguistics (ACL). Co-located with ACL-IJCNLP 2021. 1-6 August 2021.

Abeer ALDayel and Walid Magdy. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597, 2021. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2021.102597. URL https://www.sciencedirect.com/science/article/pii/S0306457321000960.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*, 2021.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

Emily M. Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6, 2011.

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.

Audrey Berquand, Paul Darm, and Annalisa Riccardi. Space transformers: language modeling for space systems. *IEEE Access*, 9:133111–133122, September 2021. ISSN 2169-3536. doi: 10.1109/ACCESS.2021.3115659.

Aivars Berzins, Khalid Choukri, Maria Giagkou, Andrea Lösch, Helene Mazo, Stelios Piperidis, Mickaël Rigault, Eileen Schnur, Lilli Small, Josef van Genabith, Andrejs Vasiljevs, Andero Adamson, Dimitra Anastasiou, Natassa Avraamides-Haratsi, Núria Bel, Zoltán Bódi, António Branco, Gerhard Budin, Virginijus Dadurkevicius, Stijn de Smeytere, Hrístina Dobreva, Rickard Domeij, Jane Dunne, Kristine Eide, Claudia Foti, Maria Gavriilidou, Thibault Grouas, Normund Gruzitis, Jan Hajic, Barbara Heinisch, Verónique Hoste, Arne Jönsson, Fryni Kakoyianni-Doa, Sabine Kirchmeier, Svetla Koeva, Lucia Konturová, Jürgen Kotzian, Simon Krek, Gauti Kristmannsson, Kaisamari Kuhmonen, Krister Lindén, Teresa Lynn, Armands Magone, Maite Melero, Laura Mihailescu, Simonetta Montemagni, Micheál Õ Conaire, Jan Odijk, Maciej Ogrodniczuk, Pavel Pecina, Jon Arild Olsen, Bolette Sandford Pedersen, David Perez, Andras Repar, Ayla Rigouts Terryn, Eirikur Rögnvaldsson, Mike Rosner, Nancy Routzouni, Claudia Soria, Alexandra Soska, Donatienne Spiteri, Marko Tadic, Carole Tiberius, Dan Tufis, Andrius Utka, Paolo Vale, Piet van den Berg, Tamás Váradi, Kadri Vare, Andreas Witt, Francois Yvon, Janis Ziedins, and Miroslav Zumrik. *Sustainable Language Data Sharing to Support Language Equality in Multilingual Europe - Why Language Data Matters: ELRC White Paper*. ELRC Consortium, 2 edition, 2019. ISBN 978-3-943853-05-6.

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.

Samuel Bowman and George Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

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.

Xieling Chen, Di Zou, Haoran Xie, and Gary Cheng. Twenty years of personalized language learning: Topic modeling and knowledge mapping. *Educational Technology & Society*, 24(1):205–222, 2021. ISSN 11763647, 14364522. URL https://www.jstor.org/stable/26977868.

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, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1017. URL https://aclanthology.org/P15-1017.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Kenneth Church, Mark Liberman, and Valia Kordoni. Benchmarking: Past, present and future. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 1–7, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.bppf-1.1. URL https://aclanthology.org/2021.bppf-1.1.

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.

Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. Combining pre-trained language models and structured knowledge. *arXiv preprint arXiv:2101.12294*, 2021.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c04c19c2c2474dbf5f7ac4372c5b9af1-Paper.pdf.

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, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Keith Cortis, Judie Attard, and Donatienne Spiteri. Malta national language technology platform: A vision for enhancing Malta's official languages using machine translation. In *Proceedings of the First Workshop on Multimodal Machine Translation for Low Resource Languages (MMTLRL 2021)*, pages 12–19, Online (Virtual Mode), September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.mmtlrl-1.3.

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.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning, 2021.

Pedro Domingos. A few useful things to know about machine learning. *Commun. ACM*, 55(10):78–87, October 2012. ISSN 0001-0782. doi: 10.1145/2347736.2347755.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

William D Eggers, Neha Malik, and Matt Gracie. Using AI to Unleash the Power of Unstructured Government Data. *Deloitte Insights*, 2019. URL https://www2.deloitte.com/content/dam/Deloitte/lu/Documents/public-sector/lu-ai-unstructured-government-data.pdf.

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.

Stephen Emmott and Anthony Mullen. Magic Quadrant for Insight Engines, 2021.

Carla Parra Escartín, Teresa Lynn, J. Moorkens, and Jane Dunne. Towards transparency in nlp shared tasks. *ArXiv*, abs/2105.05020, 2021.

European Parliament. Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI). http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf, 2018.

Boris Evelson, Srividya Sridharan, and Aldila Yunus. The Forrester Wave: AI-Based Text Analytics Platforms (Document Focused), Q2 2020, 2020.

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

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.

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.

Jose Manuel Gomez-Perez and Raúl Ortega. ISAAQ - mastering textbook questions with pre-trained transformers and bottom-up and top-down attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5469–5479, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.441. URL https://aclanthology.org/2020.emnlp-main.441.

José Manuél Gómez-Pérez, Ronald Denaux, and Andrés García-Silva. *A Practical Guide to Hybrid Natural Language Processing - Combining Neural Models and Knowledge Graphs for NLP*. Springer, 2020. ISBN 978-3-030-44829-5. doi: 10.1007/978-3-030-44830-1. URL https://doi.org/10.1007/978-3-030-44830-1.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL https://aclanthology.org/N19-1061.

Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7389–7395, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.659. URL https://aclanthology.org/2020.acl-main.659.

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.

Eran Hirsch, Alon Eirew, Ori Shapira, Avi Caciularu, Arie Cattan, Ori Ernst, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, and Ido Dagan. iFacetSum: Coreference-based interactive faceted summarization for multi-document exploration. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 283–297, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-demo.33. URL https://aclanthology.org/2021.emnlp-demo.33.

Pascal Hitzler, Federico Bianchi, Monireh Ebrahimi, and Md. Kamruzzaman Sarker. Neural-symbolic integration and the semantic web a position paper. 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Ronghang Hu and Amanpreet Singh. Transformer is all you need: Multimodal multitask learning with a unified transformer. *ArXiv*, abs/2102.10772, 2021.

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.

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, 10 2017. ISSN 2307-387X. doi: 10.1162/tacl_a_00065. URL https://doi.org/10.1162/tacl_a_00065.

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.

Divyansh Kaushik, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih. On the efficacy of adversarial data collection for question answering: Results from a large-scale randomized study. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6618–6633. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.517. URL https://doi.org/10.18653/v1/2021.acl-long.517.

Hiroaki Kitano. Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery. *AI Magazine*, 37(1):39–49, Apr. 2016. doi: 10.1609/aimag.v37i1.2642. URL https://ojs.aaai.org/index.php/aimagazine/article/view/2642.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

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.

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.

Viktorija Leonova. Review of non-english corpora annotated for emotion classification in text. In *DB&IS*, 2020.

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.

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, 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, 2021. doi: 10.24963/ijcai.2021/612. URL https://doi.org/10.24963/ijcai.2021/612. Survey Track.

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. *ArXiv*, abs/2107.13586, 2021.

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.

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.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Unicorn on rainbow: A universal commonsense reasoning model on a new multitask benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13480–13488, May 2021.

Yinquan Lu, Haonan Lu, Guirong Fu, and Qun Liu. Kelm: Knowledge enhanced pre-trained language representations with message passing on hierarchical relational graphs. *arXiv preprint arXiv:2109.04223*, 2021.

Kaixin Ma, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari. Knowledge-driven data construction for zero-shot evaluation in commonsense question answering. In *35th AAAI Conference on Artificial Intelligence*, 2021.

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.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*, 2021a.

Bonan Min, Hayley H. Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ArXiv*, abs/2111.01243, 2021b.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021c. URL https://arxiv.org/abs/2110.15943.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, 2016.

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.

Philippe Palanque and Fabio Paternò. Interactive systems: Design, specification, and verification, 7th international workshop dsv-is, limerick, ireland, june 5-6, 2000, proceedings. 01 2000. doi: 10.1109/ICSE.2000.870518.

Matthew E. Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations. 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 43–54, Hong Kong, China, November 2019a. Association for Computational Linguistics. doi: 10.18653/v1/D19-1005. URL https://aclanthology.org/D19-1005.

Matthew E Peters, Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A Smith. Knowledge enhanced contextual word representations. 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 43–54, 2019b.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online, 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.7.

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.

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.

Sarah Masud Preum, Sirajum Munir, Meiyi Ma, Mohammad Samin Yasar, David J. Stone, Ronald Williams, Homa Alemzadeh, and John A. Stankovic. A review of cognitive assistants for healthcare: Trends, prospects, and future directions. *ACM Comput. Surv.*, 53(6), feb 2021. ISSN 0360-0300. doi: 10.1145/3419368. URL https://doi.org/10.1145/3419368.

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

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

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, 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.

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.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? *CoRR*, abs/1902.10811, 2019. URL http://arxiv.org/abs/1902.10811.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Georg Rehm, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Khalid Choukri, Andrejs Vasiļjevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadina, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon. The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3315–3325, Marseille, France, 5 2020. European Language Resources Association (ELRA).

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. pages 4902–4912. Association for Computational Linguistics, 5 2020. doi: 10.18653/v1/2020.acl-main.442. URL https://www.aclweb.org/anthology/2020.acl-main.442.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1609. URL https://aclanthology.org/W17-1609.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (05):8732–8740, 2020.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*, 2021. URL https://arxiv.org/abs/2110.08207.

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.

Iulian Serban, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau. A survey of available corpora for building data-driven dialogue systems. *ArXiv*, abs/1512.05742, 2018.

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.

Ori Shapira, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan. Extending multi-document summarization evaluation to the interactive setting. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 657–677, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.54. URL https://aclanthology.org/2021.naacl-main.54.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *EMNLP (short)*, 2019.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. Societal biases in language generation: Progress and challenges. In *Proceedings of the Conference of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

Amit Sheth, Sujan Perera, Sanjaya Wijeratne, and Krishnaprasad Thirunarayan. Knowledge will propel machine understanding of content: extrapolating from current examples. In *Proceedings of the International Conference on Web Intelligence*, WI '17, pages 1–9, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4951-2. doi: 10.1145/3106426.3109448.

Yoav Shoham. Why knowledge representation matters. *Commun. ACM*, 59(1):47–49, December 2015. ISSN 0001-0782. doi: 10.1145/2803170.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2020.

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, 2019.

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.

Emma Strubell, Ananya Ganesh, and Andrew Mccallum. Energy and policy considerations for deep learning in nlp. pages 3645–3650, 01 2019. doi: 10.18653/v1/P19-1355.

Oyvind Tafjord and Peter Clark. General-purpose question-answering with macaw. *arXiv preprint arXiv:2109.02593*, 2021.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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 4149–4158, 2019.

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.

Yi Tay, Mostafa Dehghani, Jai Prakash Gupta, Vamsi Aribandi, Dara Bahri, Zhen Qin, and Donald Metzler. Are pretrained convolutions better than pretrained transformers? 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 4349–4359, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.335. URL https://aclanthology.org/2021.acl-long.335.

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.

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.

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.

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.

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, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.121. URL https://aclanthology.org/2021.findings-acl.121.

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, 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.121. URL https://aclanthology.org/2021.findings-acl.121.

Zijie J Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. Putting humans in the natural language processing loop: A survey. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 47–52, 2021c.

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. *arXiv preprint arXiv:2109.01652*, 2021. URL https://arxiv.org/abs/2109.01652.

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.

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.

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.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.572.

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.