Chapter 42
Deep Dive Text Analytics and Natural Language Understanding

Jose Manuel Gómez-Pérez, Andrés García-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiļjevs, and Teresa Lynn

Abstract In this chapter, we present a comprehensive overview of text analytics and Natural Language Understanding (NLU) from the perspective of digital language equality (DLE) in Europe. We focus on the research that is currently being undertaken in foundational methods and techniques related to these technologies as well as on the gaps that need to be addressed in order to offer improved text analytics and NLU support across languages. Our analysis includes eight recommendations that address central topics for text analytics and NLU, e. g., 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 language models and the availability of cross-lingual, cross-modal, and cross-sector datasets and benchmarks.[1]

1 Introduction

Text analytics tools have been in the market for a long time and have proven useful for extracting 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 informa-

Jose Manuel Gómez-Pérez · Andrés García-Silva · Cristian Berrio
Expert.AI, Spain, jmgomez@expert.ai, agarcia@expert.ai, cberrio@expert.ai

German Rigau · Aitor Soroa
University of the Basque Country, Spain, german.rigau@ehu.eus, a.soroa@ehu.eus

Christian Lieske · Johannes Hoffart · Felix Sasaki · Daniel Dahlmeier
SAP SE, Germany, christian.lieske@sap.com, johannes.hoffart@sap.com, felix.sasaki@sap.com, daniel.dahlmeier@sap.com

Inguna Skadiņa · Aivars Bērziņš · Andrejs Vasiļjevs
Tilde, Latvia, inguna.skadina@tilde.com, aivars.berzins@tilde.com, andrejs.vasiljevs@tilde.com

Teresa Lynn
Dublin City University, ADAPT Centre, Ireland, teresa.lynn@adaptcentre.ie

[1] This chapter is an abridged version of Gomez-Perez et al. (2022).

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_42

tion contained in the text. Typically, such a process starts by extracting relevant data from text that is later used in analytics engines to derive additional insights. Nowadays text analysts have a wide range of accurate features available to help recognise and explore patterns when 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 (NLP). NLP 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 NLP frequently involve natural language understanding (NLU), natural language generation, connecting language and machine perception, dialogue systems, and their combination.

Recent breakthroughs in deep learning have resulted in impressive progress in NLP. Neural language models like BERT and GPT-3 are able to infer linguistic knowledge from large collections of text that can then be transferred to deal effectively with NLP tasks without requiring too much additional effort. Neural language models have had a positive impact on key tasks of text analytics and NLU, 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 has embraced this technology and hybrid tools are emerging nowadays, combining or replacing robust rule-based systems that used to be the norm in the market with machine learning methods. Nevertheless, despite all the hype about data-driven approaches to text processing and particularly Transformer-based language models like BERT (Devlin et al. 2019), which might lead non-experts to think that everything is already solved in text analysis and NLU, many gaps still need to be addressed to make state-of-the-art language technologies (LTs) fully operational and benefit all European languages. Especially relevant is the fact that data-driven approaches require very large amounts of data for training.

Language models have lessened the requirement of labelled data to address downstream tasks, but the need for such data has not disappeared. Beyond general purpose datasets, labelled data is scarce, labour-intensive and thus expensive to produce. Access to labelled data is one of the major hurdles in leveraging data-driven approaches in business applications, and is especially problematic for under-resourced languages for which such data does not exist in sufficient quantities, 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 explanatory capability is a major obstacle to bringing such technology in domains where regulation demands systems which can justify every decision they make. Furthermore, language models pose ethical challenges including gender and racial biases that are learned 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 NLU tools that provide appropriate coverage for all European languages, particularly for minority and regional languages. Thus, one of the main goals of this chapter is to outline how the European text analytics industry and research community can address the shortcomings by building on the strengths of current text analytics and NLU tools. We call for human-centric text analysis where people's knowledge, emotions and needs are put at the centre of the design and learning process of the next generation of tools. Other topics in the research agenda are hybrid approaches combining existing rule-based and data-driven systems, multilingualism in text analytics, multimodal analysis of information, and a new generation of benchmarks.

1.1 Scope of this Deep Dive

To better understand how text analytics and NLU 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 offerings of the main text analytics and NLU providers in the European market. This analysis also includes recent findings in related research areas, such as NLP/NLU, machine learning, and information retrieval, where language understanding tasks that not long ago were the 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, categorisation, and topic modelling, 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 specific domains where such 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 and employee experience, brand management, recruiting, or contract analysis. An exhaustive account of each sector and use-case, and their relevance for text analytics, is out of scope of this chapter.

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 text analytics tools. This chapter addresses language coverage as another key dimension for the analysis of text analytics and NLU tools when considering DLE.

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. However, more sophisticated approaches are needed as we transition towards scenarios involving a deeper understanding of text in order to solve increasingly complex tasks like abstractive summarisation, reading comprehension, recognising textual entailment, or stance detection. Therefore, this chapter puts a 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 algorithms 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 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. Another important area of research that we cover deals with leveraging other modalities of information in addition to text.

All such aspects are considered from the perspective of their combined impact on society. We provide recommendations to address the current limitations of text analytics and NLU technologies in the interest of promoting DLE in Europe.

1.2 Main Components

The goal of text analytics is to discover novel and interesting information in documents and text collections that can be, among others, useful for further analysis or strategic decision-making. Text analytics tools support a wide range of functionalities to process, leverage and curate texts. Most of these functionalities can be broadly categorised into syntactic 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 that requires machine-reading comprehension, has made the transition from research labs to production systems.

The challenges involved in NLP and NLU have different levels of complexity, and as a result, the solution to each of the many challenges is at a different level of progress. For example, natural language generation is one such challenge, where recent advances like GPT-3 are heralded as a key enabler for a new generation of language applications.[2] Therefore, in addition to functionalities that are already available in the market, there are others which the research community is currently working on. Some advanced functionalities involve reasoning, 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 have arisen to address hard problems involving text generation, e. g., *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 from data contained in tables or JSON documents.

Recently, commercial text analytics providers have started supporting the customisation of functionalities, e. g., users can define classes, entity and relation types, or sentiment scores. This is possible thanks to supervised machine learning making use of user-generated examples. The user only provides examples while the text analytics tool handles all the complexity of the machine learning process. Thus, end users do not need a background in ML to customise 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, but 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.

To study the language support of existing text analytics technologies and NLU tools, we look in two main directions: 1. the catalogue of services of global technology providers, which provides us with a notion of what is being currently made available and marketed to the public; and 2. European initiatives that offer repositories of language resources and tools (LRTs), like the European Language Grid (ELG, Rehm 2023). At the time of writing, the ELG catalogue holds more than 11,500 metadata records (Labropoulou et al. 2020), including both data and tools/services, covering almost all European languages.[3] The ELG platform was populated with more than 6,000 additional language resources identified by language informants in the ELE consortium and harvests major EU LRT repositories such as CLARIN[4] and ELRC-SHARE.[5] The observations and figures included in this chapter have been extracted from ELG, which aims at concentrating all available resources, tools and services and making them available in a single platform. Our goal with this chapter is not

[2] https://openai.com/blog/gpt-3-apps/

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

[4] https://www.clarin.eu

[5] https://elrc-share.eu

to provide an exhaustive account, for which such figures could be complemented with additional information from other European infrastructures like the ones mentioned above, but rather to provide an up-to-date indication of the support that each European (and non-European) language enjoys.

For commercial text analytics services, we draw on reports from key players in market intelligence such as Gartner Magic Quadrant for Insight Engines[6] and the Forrester Wave: AI-Based Text Analytics Platforms 2020.[7] A mandatory requirement for providers to be included in this study is for service documentation be 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 other recognised providers, like 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 European 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 immigrants in Europe; 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

A summary of our findings follows. A small set of services including entity extraction, key phrase extraction, and syntactic analysis, offered by global text analytics providers, 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 European languages in category B is actually the worst: the language support of the functional services is scarce or non-existent. Languages in category C also have low coverage across all functional services. In contrast, custom entity extraction has almost perfect support of the languages across all categories. However, custom classification, custom sentiment analysis, and custom relation extraction have a language coverage similar to off-the-shelf text analytics services, covering less than half of the languages in categories A and C, and barely any language at all in category B.

According to the ELG catalogue, syntactic analysis services (language identification, tokenization, etc.) are available for nearly all languages in category A. However, the language support of such services drops to 63% of languages in category B, and 72% in category C. Named entity recognition has moderate support across all language categories reaching 66% for category A, 54% for category B and 61%

[6] https://www.gartner.com/en/documents/3999454

[7] https://www.forrester.com/report/The-Forrester-Wave-AIBased-Text-Analytics-Platforms-Document-Focused-Q2-2020/RES159887

for category C. From there, language support for text analytics services such as keyword extraction, sentiment analysis, summarisation, and entity linking is poor or non-existent in every language category.

Our analysis shows that official EU languages are covered by a subset of text analytics services including syntactic analysis, key phrase extraction, and entity extraction. However, only a small fraction of category A languages are supported by the remaining services. For other European languages in category B, global players offer scarce support or none at all, and for languages in category C support is also low. In ELG the picture changes a little for category B languages since the number of supported languages increases for some of the functional services. However, overall support of languages in categories B and C is still low, i. e., global players plan their offerings based on the volume of the potential market for each language.

2 State-of-the-Art and Main Gaps

2.1 State-of-the-Art

LRTs have increased and improved since the end of the 1990s, a process further catalysed by the advent of deep learning and neural networks and lately with large pre-trained language models. Today, NLP practitioners find themselves in the midst of a paradigm shift. This revolution has brought noteworthy advances to the field. 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. 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 in terms of digital presence. Moreover, the striking asymmetry between official and non-official European languages with respect to available digital resources is worrisome.

Unfortunately, European DLE is failing to keep pace with these rapidly evolving changes. Neural language models and related techniques are key to NLP progress and so being able to build them for target languages with the same quality as English is key if language equality is to be achieved. Now is the moment to seek balance between European languages in the digital realm. There are ample reasons for optimism. Although there is more work that can and must be done, Europe’s leading LRT repositories, platforms, libraries, models and benchmarks have begun to make inroads. Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pre-trained language models and self-supervised systems opens up the way to leverage NLP for less developed languages.

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 different factors: 1. mature deep learning technology; 2. large amounts of data (including multilingual text data); 3. increase in HPC (GPUs); 4. application of simple but effective self-learning and transfer learning approaches using Transformers. 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. 2021).

2.2 Main Gaps

We focus on eight main areas related to text analytics and NLU that have an impact on digital language equality: data, legal aspects, limitations, benchmarking, conformance, and domain experts' tooling.

Data – The availability of suitable data for training and evaluating NLP tools is crucial. Unfortunately, current language data for text analytics suffers from several shortcomings. Labelling data can be a lengthy operation that requires skilled domain expertise, which is costly and hard to find. Data and language coverage is a concerning issue as the majority of datasets that are relevant to Europe are general-purpose datasets based on major languages such as English, German, Spanish and French. However, under the EU Digital Europe Programme, new common Data Spaces, including a Language Data Space, will be created. Quality is also important: reliable (misinformation-free), balanced (no bias) and clean content (non-toxic/hate-speech). Machine learning models are notoriously sensitive to bias and noise within datasets. Thus, there is a clear need for reliable bias and toxicity detection tools.

Legal aspects – Since text can 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 important for EU citizens' protection, significantly hampers language data sourcing and reuse 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 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.

NLU limitations – Most of today's text analytics solutions are language-specific. 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. As we have seen, data availability is already a general problem, but when it comes to lesser-spoken languages with lower amounts of digital content, such scarcity is compounded. Similarly, key pieces of contextual information such as the author, intended audience, societal factors and the purpose of communication also need to 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 to make better judgements or predictions. One priority for many businesses and organisations is to build trust and confidence in 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. Finally, further exploration is required into extensibility methods to include domain-specific knowledge (e. g., when large corpora are not available), allowing LT providers to easily build custom extensions for machine learning-based systems.

Benchmarking – In language technology (and NLU in particular), a wide range of benchmarking frameworks exists 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 or, say, SacreBLEU[8] scores for dialogue systems. Current NLU benchmarks include widely adopted ones like GLUE and and SuperGLUE.[9] In terms of the nature of datasets used in benchmarking, realistic data is lacking. Therefore, the increasing trend for creating (often general purpose) synthetic data proves to be problematic. Some evaluation datasets are also often criticised in academic shared tasks, where they are sometimes referred to as 'toy' examples that are not applicable to real-world problems. There is a clear need for an increase in diversity, relevance and suitability of annotated test data.

Conformance – A dimension related to standards concerns conformance, namely "the fulfillment of specified requirements by a product, process, or service."[10] 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, requirement statements are needed for any text analytics 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[11] in Spanish and Portuguese.[12] In particular, in the context of regulated industries, certification may need to be considered.

Domain experts tooling – Today, most work in LT based on ML requires expert level skills in tools related to data management, data science and NLP. This creates bottlenecks since it does not allow domain experts (e. g., experts in finance) to become actively involved without extensive tool training or understanding of the underlying technology. This setup causes overhead and delays since work between tool experts and domain experts needs to be coordinated. What is lacking as a way to address this is the availability of consumer-grade, highly usable, low code or no code tools for domain experts. Ideally, such tools should be developed in collaboration with usability specialists, to allow domain experts to play a more active role in the development of solutions for application scenarios they are familiar with.

[8] https://huggingface.co/metrics/sacrebleu

[9] https://gluebenchmark.com, https://super.gluebenchmark.com

[10] https://www.w3.org/TR/qaframe-spec/#specifying-conformance

[11] https://cldr.unicode.org/index/downloads

[12] 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.

3 The Future of the Area

3.1 Contribution to Digital Language Equality

Today, text analytics tools can help societies and individuals in various ways by supporting tasks that involve the discovery of information (facts, rules, relationships) in text. 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 analytics tools are also widely used in online and social media data analysis of use to both businesses and governments.

Currently, however, all of these advances and digital innovations are really only supporting major well-resourced languages (i. e., 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, applications need to be built upon systems that understand the underlying patterns in each language that requires support. As today's NLP techniques are increasingly data-driven, this means that sufficient amounts of data need to be made available in order to adapt technologies to these languages. However, even here, it may not be as simple as plugging in new datasets to existing technologies; due to the fact that languages and domains can differ so significantly, various types of parameter tuning, system adaptation or hybrid implementation may also be required to achieve robust and reliable technologies in new languages and scenarios.

Text analytics 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. 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, including 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 analytics tools (together with machine translation solutions and other cross- and multi-lingual solutions) are key for accessing this information and knowledge across Europe.

The META-NET White Paper Series (Rehm and Uszkoreit 2012) reported on an analysis of LRTs available for EU languages. The results showed that with respect to text analytics, *good support* only applied to English, and *moderate support* to five widely spoken languages: Dutch, French, German, Italian and Spanish. This meant that the other 24 (out of 30) European languages in this study were clustered under *fragmented* as well as *weak or no support*. Today, all 24 official EU languages benefit from basic tools: tokenizers, lemmatizers, morphological analysers, part-of-speech tagging tools, and syntactic parsers. While the quality, reliability or robustness of these tools vary across languages, their existence represents a step in the right direction. In contrast, more sophisticated tools and services (e. g., summarisation tools) are available only for a small number of languages.

Some of the main reasons that prevent sophisticated text analytics techniques from being 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 text analytics technologies, and the complexity of technology adaptation in low-resource settings. For instance, in the case of dialogue systems and chatbots, analysis of available datasets for dialogue modelling clearly demonstrates a gap for less-resourced languages (Serban et al. 2018; Leonova 2020).

Gartner (2021) forecasts the worldwide AI software revenue to $62.5 billion in 2022, an increase of 21.3% from 2021. Intelligent, AI-based, virtual assistants are already in demand in the digital market and their use in the workplace is growing. Gartner (2020) predicts that by 2025, 50% of knowledge workers will use a virtual assistant on a daily basis, up from 2% in 2019. For the public sector and businesses, this provides an opportunity to use intelligent virtual assistant technology to take care of more repetitive and auxiliary business processes. Gartner (2019) predicts that decision support/augmentation will be the largest area of AI by 2030, accounting for 44% of business value, with agents representing 24%.

For countries with lesser-spoken languages, these predictions only hold if technology exists to support them, of course. If not, an economic divide will emerge, as countries with sufficient language technologies will gain (further) advantage.

3.2 Breakthroughs Needed

Various global enterprises from the US and Asia have started deploying large pre-trained neural language models in production. However, despite their impressive capabilities, large language models raise severe concerns. Currently, we have no clear understanding of how they work, when they fail, and which emergent properties they present. As argued by Bender et al. (2021), it is important to understand the limitations of language models, which they call “stochastic parrots”, and put their success in perspective. 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 harmful stereotypes, and to the inclusion of personal information. Moreover, these models are costly to train and develop, both financially and environmentally. 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 them (Ahmed and Wahed 2020).

To tackle these questions, much more critical interdisciplinary collaboration and research are needed. In Europe there is a lack of necessary resources (experts, data, computing facilities, etc.) compared to large US and Chinese IT enterprises that lead the development of these new 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 democratising 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. We note with concern a tendency to focus on state-of-the-art results exclusively with the help of leaderboards, without encouraging a deeper understanding of the mechanisms by which they are achieved. We believe that such short-term goals can generate misleading conclusions and direct resources away from important efforts that facilitate long-term progress towards efficient, accurate, explainable, ethical and unbiased multilingual language understanding. Progress in these fields will help achieve DLE in Europe in all aspects of society, from government to businesses to the citizens themselves. Next, we focus on some of these key technical areas.

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. 2022). *Prompting* is a technique that involves adding a piece of text (prompt) to the input examples to "encourage" a language model to bring to the surface the implicit knowledge the user is interested in, i. e., guiding 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. 2022; Tafjord and Clark 2021). The application of zero-shot to few-shot transfer learning with multilingual pre-trained language models, prompt learning, and self-supervised systems opens up opportunities for less developed languages in NLP.

Integrating common sense knowledge and reasoning in NLP systems has traditionally been seen as a nearly impossible goal. 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. 2021; 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 pre-trained neural language models. This motivates new, relatively under-explored research avenues in common sense 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 ways to leverage structured and unstructured information sources and to enhance contextual representations with structured, human-curated knowledge (Peters et al. 2019; Colon-Hernandez et al. 2021; Lu et al. 2021). Despite perhaps overly optimistic claims of human parity in many tasks, *Natural Language Understanding 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 it is required to properly understand natural language (Bender and Koller 2020).

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, is still far from perfect. For example, the textual snippets generated by advanced language models such as GPT and successors are formed by syntactically correct sentences that seem to talk about a particular topic, however, there is often a lack of coherence among them and humans still need to monitor and adapt the output of such systems. There is a growing body of research of human-in-the-loop NLP frameworks, where model developers continuously integrate human feedback into the model deployment workflow. These feedback loops cultivate a human-AI partnership that enhances model accuracy and robustness and builds users' trust in NLP systems (Z. J. Wang et al. 2021). In the foreseeable future we expect more such interactions, as AI and NLP become embedded in everyday work processes.

While the NLP community is fully committed to the open-source culture, the aspect of reproducibility has been less of a concern, although the topic is becoming a central one in NLP. Nowadays the majority of scientific articles are accompanied by the source code and data required to reproduce the experiments. Leaderboards such as NLP-progress,[13] Allen Institute of AI leaderboard,[14] Papers with code,[15] or Kaggle[16] encourage participation and facilitate evaluation across many different tasks and datasets. As a result, the NLP community has considerably increased access to publicly available and easily accessible models and datasets. 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. Open-source libraries such as Transformers[17] may open up these advances to a wider LT community. This library consists of carefully engineered state-of-the art Transformer architectures under a unified API and a curated collection of models (Wolf et al. 2020a). Following up on the success of the Hugging Face platform (Wolf et al. 2020b), 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.[18]

3.3 Technology Visions and Development Goals

In this section, we provide an overview of the main technological visions for NLP and NLU, which will contribute to achieving DLE in Europe by 2030. We have identified developments for increasing the language support of such technologies, putting

[13] http://nlpprogress.com

[14] https://leaderboard.allenai.org

[15] https://paperswithcode.com/area/natural-language-processing

[16] https://www.kaggle.com/datasets?tags=13204-NLP

[17] https://huggingface.co

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

users' needs at the centre of any breakthroughs involving language technologies, the integration with other modalities of information in addition to text, the hybridisation of symbolic AI and neural systems, and the need for a new benchmarking approach.

Language support beyond widely spoken languages, including minority and under-resourced languages, is still a pending issue in text analytics and NLU. The investment of LT providers in such languages is inhibited most probably due to a comparatively lower profitability in this space compared to mainstream languages, considering the number of potential users. Nevertheless, the current trend in LT relying on neural language models and research on unsupervised and zero-shot learning opens up 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 settings in a wide range of tasks (Radford et al. 2019; Brown et al. 2020; Gao et al. 2021). This is primarily due to the fact that language models learn to perform tasks from patterns occurring in text, eliminating or reducing to a great extent the need for additional labeled data which is a scarce resource for many languages.

Despite their dominance in current NLP pipelines, language models have mainly been addressed as a one-size-fits-all approach, offering almost no customisation beyond the data used to fine-tune (Devlin et al. 2019) or prompt (Brown et al. 2020) models for downstream tasks. Current research focused on unsupervised and zero-shot learning (Gao et al. 2021) delves into this issue since users have little to say in the learning process. Moreover, the data-driven approach and race for accuracy have yielded opaque tools that are hard to interpret, and biased tools that perpetuate social stereotypes related to gender, race and ethnicity in text collections. The lack of transparency makes it difficult to build trust between users and system predictions, having negative consequences for technology adoption. Biased tools have a direct impact on society, especially for marginalised populations (Sheng et al. 2021).

We advocate for a *next generation of language tools that care about end user needs and expectations*, making them part of the design and learning process. These tools will be human-aware, encompass human emotions, and be trustworthy, avoid bias, offer explanations, and respect user privacy. Moreover, human intelligence will be used together with machine learning techniques to produce better LRTs. Human feedback will be a guide in the learning process, informing the machine as to what users want or do not want. Reinforcement learning from human feedback is a promising research avenue (Stiennon et al. 2020; Li et al. 2016) to use human intelligence to improve NLP tools. Also, interactivity with domain experts and users (e. g., Shapira et al. 2021) is a key area for further advances beyond the usual supervised paradigm.

As practitioners come to realise the inevitable limitations of purely end-to-end deep learning approaches, which increase in the case of under-represented languages (both in terms of available language models and suitable training corpora), the *transition to hybrid approaches involving different ways of combining neural and symbolic approaches* 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 interact, and how we should evaluate the resulting systems using appropriate benchmarks. The field of neurosymbolic approaches 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. 2019) and K-Adapter (R. Wang et al. 2021), not only to make NLU models aware of the entities contained in a knowledge base and the relations between them from a general point of view, as provided by resources like Wikipedia or Wikidata, but also when it comes to quickly incorporating existing resources from vertical domains and custom organisations into our models in a fast, scalable way. Some, e. g., Sheth et al. (2017) and Shoham (2015), argue that knowledge graphs can enhance both expressivity and reasoning power in machine learning architectures. Others (Gómez-Pérez et al. 2020) propose a working methodology[19] 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.

Different modalities can be combined to provide complementary information that may be redundant but can help to convey information more effectively (Palanque and Paternò 2000). For example, 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 have been conducted individually such as NLP, automatic speech recognition and computer vision. Deep learning techniques will play an important role in multimodal analysis. Recently, Transformer architectures (Devlin et al. 2019), initially proposed for NLP, have been used for image processing (Dosovitskiy et al. 2021) and cross-modal information processing including images and text (Hu and Singh 2021). Other approaches based on contrastive language-image pre-training, like CLIP (Radford et al. 2021), emphasise the relevance of 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. Unfortunately, such text is in English only, showing how language inequality also impacts language-vision tasks.

Benchmarking aligns research with development, engineering with marketing, and competitors across the industry in pursuit of a clear objective. However, for many NLU tasks evaluation 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 benchmarks in favour of adversarially constructed, out-of-distribution test sets ensures that current models will perform poorly, but ultimately only serves to obscure the abilities that we want our benchmarks to measure. Adversarial data collection, understood as the process in which a human workforce interacts with a model in real time, attempts to produce examples that elicit incorrect predictions, but does not meaningfully address the causes of model failures, as shown, for instance, by Kaushik et al. (2021) for question answering. Restoring a healthy evaluation ecosystem will require significant progress

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

in the design of benchmark datasets, the reliability with which they are annotated, their size, and ways in which they handle social bias. This is even more important when we expand our view to the multilingual landscape, such as the European multilingual reality. Furthermore, much more emphasis will need to be given to typical realistic settings (Church et al. 2021), 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 illustrated by Rajpurkar et al. (2018) inter alia. Recht et al. (2019) also showed that this can lead to overestimating real-world performance. Approaches like Ribeiro et al. (2020) advocate for a methodology that breaks down potential capability failures into specific behaviours, 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. 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. In contrast, such benchmarks need to be multilingual by design and cover each economic sector for each of the European languages, guaranteeing language equality regardless of the size of the market associated with each language.

3.4 Towards Deep Natural Language Understanding

Much has been said about the impact of intelligent systems on our lives. Today's large amounts of available data, produced at an increasing pace and in heterogeneous formats and modalities, have stimulated the development of means that extend human cognitive and decision-making capabilities, alleviating such burdens and assisting our drivers, doctors, teachers and scientists. In scientific disciplines like biomedical sciences, some like Kitano (2016) even propose a new grand challenge for this kind of systems: to develop an AI that can make major scientific discoveries that are eventually worthy of a Nobel Prize. This suggests the time is ripe for a shared partnership with machines, where humans can benefit from augmented reasoning and information management capabilities. Through such a partnership, we foresee a virtuous circle of data collection, active learning, and interactive feedback, which will result in adaptive, ever-learning systems.

We have already seen signs of such a partnership, e. g., in the application of generative models like GPT-3 to produce text given a prompt, with applications in different business sectors. Based on these developments, some suggest[20] that the future of AI lies in the development of systems that allow maintaining a conversation with a computer. This scenario should go beyond current and past chatbots, able to copy form without understanding meaning but nevertheless capable of creating a dialogue

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

with the user. However, this often seems to be missing from AI systems like facial recognition algorithms, which are imposed upon us, or self-driving cars, where the public becomes the test subjects in a potentially dangerous experiment. Language will require advances in knowledge representation, true understanding of meaning and pragmatics, and the ability of models to explain and interpret their predictions in ways that humans can understand and relate to.

The AI community and particularly the areas related to text understanding also need to address issues like fairness in ways that tangibly and directly benefit disadvantaged and misrepresented populations. We have spent large amounts of effort discussing fairness and transparency in our algorithms. At the algorithmic level, fairness has to do with the absence of bias in the models that for example in NLU are used to address tasks that may range from the evaluation of mortgage applications or insurance policies to medical examinations and career recommendations. If algorithms are biased, so are their predictions, in which case inequalities would be perpetuated as AI technologies are deployed more and more in society.

This is essential work. The lack of resources in a specific language to train an NLU model in that language can be seen as 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 apps pose major challenges and risks if users cannot customise suggested routes in order to avoid insurmountable obstacles. Similarly, the lack of availability of service functionalities in all languages will have an unwanted effect in the respective populations. Accessibility, education, homelessness, human trafficking, misinformation, and health among others are all areas where AI and text understanding can have a really positive impact on people's quality of life. So far, we have only started to scratch the surface.

4 Summary and Conclusions

We finish this chapter with a list of recommendations and guidelines that address central topics for text analytics and NLU. Among others, we emphasise the role of language equality for social good, the balance between commercial interests and equal opportunities for society, and incentives to help bring about language equality. We also focus on key technologies like neural language models and the availability of multilingual, cross-sectorial datasets and benchmarks.

1. *Language equality in text analytics is a transformative and integrative force for social good* that 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 will encourage progress and prosperity, creating new markets for text analytics and other areas related to AI and LT across Europe. However, this is not yet a common scenario for all European languages. The question we should ask ourselves is: what is the alternative? What will the

social cost be if the required policies do not effectively reach *all* European languages until 2030?

2. *The balance between legitimate commercial interests and equal access to opportunities is fragile* when it comes to DLE in text analytics. We have shown how global providers tend to concentrate their offerings and investment in more widespread languages, neglecting a long tail of languages with smaller populations. In contrast, European initiatives such as ELG (Rehm 2023) provide a more equitable coverage. Two reflections emerge. First, it is a European priority to ensure that all European languages are properly covered. Therefore, European companies and also European research organisations 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, but also in other forms, like – for companies – tax breaks associated to language services for less represented languages or – for research organisations – specific regional or national funding that can only be used for developing tools or resources for the national or regional language. Second, to create traction this effort should involve European technology providers but also consumers of such services at the different levels of the European public administration and large European companies.
3. *Possible incentives to language equality in text analytics and NLU are not just financial.* Acknowledging that we are working on a particular language conveys the opportunity to stress that research is language-specific. Conversely, neglecting to state that a particular piece of research worked on, say, English language data gives a false veneer of language independence (Bender 2011). Incentives need to be provided for Text Analysis research to cover *all* European languages.
4. *Neural language models are a cornerstone of most NLU and text analytics pipelines now, and this will continue in the next few years.* However, current methods to create such models are hardware-intensive, require vast amounts of text data, and the training comes at the cost of high energy consumption and a large carbon footprint. Because of this, most of the 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.
5. *Data is key.* Without sufficient amounts of good-quality data, language models and text analytics solutions based on ML 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 tasks like classification, sentiment analysis, etc. While much progress has been made in creating large-scale labeled data sets for the major languages, it is not yet feasible, especially from a business-driven perspective, to do this for all European languages, let alone the literally thousands of languages spoken on the planet. As suggested in the previous item, 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 our

languages following self-supervised approaches. The problem comes in satisfying the needs of domain- and task-specific data to adapt such models to solve real-life problems in each of the different business sectors and languages.

6. *Data tends to be locked in regulatory and corporate silos.* Research and solutions for LTs that address problems of business and social relevance is underdeveloped. A major reason is that enterprise data is not available to researchers in 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 through which companies can make data available under certain terms still needs to crystallise into a dynamic 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. This effort would benefit from an improved balance between European regulations like GDPR and the use of data for research purposes. Currently, companies abiding by GDPR face restrictions and demands that impose some burdens. To be competitive, European companies may need to use neural language models built by third parties in the US or China that are not subject to such regulations.
7. *Benchmarking is inadequate and needs to be fixed and updated.* For many NLU tasks evaluation 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 favour of adversarially-constructed, out-of-distribution test sets means that current models will perform poorly, and ultimately only obscures the abilities that we want our benchmarks to measure. Restoring a healthy evaluation ecosystem, particularly one involving a vision for DLE, 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, if we want to make well-grounded progress it is crucial that improved benchmarking considers not only technical but also ethical and societal issues. Benchmark design needs to fit realistic data compositions, rather than synthetic ones within our comfort zone. Addressing such shortage of real-life benchmarks will require significant collaboration between European industry and academia.
8. *Text does not live in isolation. Information is cross-modal.* Text is rarely found in isolation in real-life. Addressing many of the market and societal challenges towards DLE will benefit from taking into account cross-modal scenarios to leverage additional sources of free supervision. Recent advances like OpenAI's CLIP and Meta's Data2Vec[21] seem promising. However, perhaps not surprisingly, all such models are currently available in English only.

[21] https://ai.facebook.com/research/data2vec-a-general-framework-for-self-supervised-learning-in-speech-vision-and-language

Finally, we would like to emphasise two points that are particularly critical to ensure DLE in Europe. First, *neural language models and related techniques are at the core* of sustaining progress in LT in modern NLP. Therefore, being able to build language models for target languages with the same quality as English is key for language equality. Second, *multilingual data is the key element* to train such models in the target languages. We should not assume that large amounts of publicly available corpora of good quality can be readily obtained for all European languages, but 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 centre of any future efforts striving for DLE.

References

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

Bender, Emily M. (2011). "On Achieving and Evaluating Language-Independence in NLP". In: *Linguistic Issues in Language Technology* 6.

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

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

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

Church, Kenneth, Mark Liberman, and Valia Kordoni (2021). "Benchmarking: Past, Present and Future". In: *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*. Online: Association for Computational Linguistics, pp. 1–7. DOI: 10.18653/v1/2021.bppf-1.1.

Colon-Hernandez, Pedro, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal (2021). "Combining Pre-Trained Language Models and Structured Knowledge". In: *arXiv preprint arXiv:2101.12294*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *NAACL Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. https://aclanthology.org/N19-1423.

Ding, Ning, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun (2022). "OpenPrompt: An Open-source Framework for Prompt-learning". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demon-*

strations. Dublin, Ireland: Association for Computational Linguistics, pp. 105–113. https://aclanthology.org/2022.acl-demo.10.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929 `[cs.CV]`.

European Parliament (2018). *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.

Gao, Tianyu, Adam Fisch, and Danqi Chen (2021). “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)*. Online: Association for Computational Linguistics, pp. 3816–3830. https://aclanthology.org/2021.acl-long.295.

Gómez-Pérez, José Manuél, Ronald Denaux, and Andrés García-Silva (2020). *A Practical Guide to Hybrid Natural Language Processing - Combining Neural Models and Knowledge Graphs for NLP*. Springer. DOI: 10.1007/978-3-030-44830-1.

Gomez-Perez, Jose Manuel, Andres Garcia-Silva, Cristian Berrio, German Rigau, Aitor Soroa, Christian Lieske, Johannes Hoffart, Felix Sasaki, Daniel Dahlmeier, Inguna Skadiņa, Aivars Bērziņš, Andrejs Vasiļjevs, and Teresa Lynn (2022). *Deliverable D2.15 Technology Deep Dive – Text Analytics, Text and Data Mining, NLU*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/text-analytics-deep-dive.pdf.

Gomez-Perez, Jose Manuel and Raúl Ortega (2020). “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)*. Online: Association for Computational Linguistics, pp. 5469–5479. https://aclanthology.org/2020.emnlp-main.441.

Han, Xu, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu (2021). “Pre-Trained Models: Past, Present and Future”. In: *AI Open* 2, pp. 225–250. https://www.sciencedirect.com/science/article/pii/S2666651021000231.

Hu, Ronghang and Amanpreet Singh (2021). “Transformer is all you need: Multimodal multitask learning with a unified transformer”. In: *arXiv preprint arXiv:2102.10772* 2.

Kaushik, Divyansh, Douwe Kiela, Zachary C. Lipton, and Wen-tau Yih (2021). “On the Efficacy of Adversarial Data Collection for Question Answering: Results from a Large-Scale Randomized Study”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 6618–6633. https://aclanthology.org/2021.acl-long.517.

Kitano, Hiroaki (2016). “Artificial Intelligence to Win the Nobel Prize and Beyond: Creating the Engine for Scientific Discovery”. In: *AI Magazine* 37, pp. 39–49. DOI: 10.1609/aimag.v37i1.2642.

Labropoulou, Penny, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva (2020). “Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid”. In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by 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. Marseille, France: ELRA, pp. 3421–3430. https://www.aclweb.org/anthology/2020.lrec-1.420/.

Leonova, Viktorija (2020). “Review of Non-English Corpora Annotated for Emotion Classification in Text”. In: *Databases and Information Systems – 14th International Baltic Conference, DB&IS 2020, Tallinn, Estonia, June 16-19, 2020, Proceedings*.

Li, Jiwei, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao (2016). “Deep Reinforcement Learning for Dialogue Generation”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 1192–1202. https://aclanthology.org/D16-1127.

Lourie, Nicholas, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi (2021). “UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.15, pp. 13480–13488.

Lu, Yinquan, Haonan Lu, Guirong Fu, and Qun Liu (2021). “KELM: Knowledge Enhanced Pre-Trained Language Representations with Message Passing on Hierarchical Relational Graphs”. In: *arXiv preprint arXiv:2109.04223*.

Ma, Kaixin, Filip Ilievski, Jonathan Francis, Yonatan Bisk, Eric Nyberg, and Alessandro Oltramari (2021). “Knowledge-Driven Data Construction for Zero-shot Evaluation in Commonsense Question Answering”. In: *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, pp. 13507–13515. https://ojs.aaai.org/index.php/AAAI/article/view/17593.

Min, Bonan, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth (2021). “Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey”. In: *arXiv preprint arXiv:2111.01243*.

Mostafazadeh, Nasrin, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen (2016). “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*. San Diego, California: Association for Computational Linguistics, pp. 839–849. https://aclanthology.org/N16-1098.

Palanque, Philippe and Fabio Paternò, eds. (2000). *Interactive Systems: Design, Specification, and Verification, 7th International Workshop DSV-IS, Limerick, Ireland, June 5-6, 2000, Proceedings*. DOI: 10.1109/ICSE.2000.870518.

Peters, Matthew E., Mark Neumann, Robert Logan, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith (2019). “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)*. Hong Kong, China: Association for Computational Linguistics, pp. 43–54. https://aclanthology.org/D19-1005.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever (2021). “Learning Transferable Visual Models From Natural Language Supervision”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. PMLR, pp. 8748–8763.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). *Language Models are Unsupervised Multitask Learners*. Tech. rep. OpenAI.

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

Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar (2019). “Do ImageNet Classifiers Generalize to ImageNet?” In: *Proceedings of the 36th International Conference on Machine Learning*. Long Beach. https://proceedings.mlr.press/v97/recht19a/recht19a.pdf.

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

Rehm, Georg, 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 Tufiş, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon (2020). "The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe". In: *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*. Ed. by 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. Marseille, France: ELRA, pp. 3315–3325. https://www.aclweb.org/anthology/2020.lrec-1.407/.

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

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh (2020). "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 4902–4912. https://www.aclweb.org/anthology/2020.acl-main.442.

Sakaguchi, Keisuke, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi (2021). "WinoGrande: An Adversarial Winograd Schema Challenge at Scale". In: *Communications of the ACM* 64.9, pp. 99–106. https://doi.org/10.1145/3474381.

Sanh, Victor, 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, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal V. Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng-Xin Yong, Harshit Pandey, Michael Mckenna, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush (2022). "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *ICLR 2022 – Tenth International Conference on Learning Representations*. Online. https://hal.inria.fr/hal-03540072.

Serban, Iulian, Ryan Lowe, Peter Henderson, Laurent Charlin, and Joelle Pineau (2018). "A Survey of Available Corpora for Building Data-Driven Dialogue Systems". In: https://arxiv.org/abs/1512.05742.

Shapira, Ori, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, Yael Amsterdamer, and Ido Dagan (2021). "Extending Multi-Document Summarization Evaluation to the Interactive Setting". In: *Proceedings of the 2021 North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online, pp. 657–677. DOI: 10.18653/v1/2021.naacl-main.54.

Sheng, Emily, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng (2021). "Societal Biases in Language Generation: Progress and Challenges". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 4275–4293. https://aclanthology.org/2021.acl-long.330.

Sheth, Amit, Sujan Perera, Sanjaya Wijeratne, and Krishnaprasad Thirunarayan (2017). "Knowledge Will Propel Machine Understanding of Content: Extrapolating from Current Examples". In: *Proceedings of the International Conference on Web Intelligence*. Leipzig, Germany: ACM, pp. 1–9. DOI: 10.1145/3106426.3109448.

Shoham, Yoav (2015). "Why Knowledge Representation Matters". In: *Communications of the ACM* 59.1, pp. 47–49. DOI: 10.1145/2803170.

Stiennon, Nisan, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F. Christiano (2020). "Learning to Summarize with Human Feedback". In: *Advances in Neural Information Processing Systems* 33, pp. 3008–3021.

Tafjord, Oyvind and Peter Clark (2021). "General-Purpose Question-Answering with Macaw". In: *ArXiv* abs/2109.02593.

Talmor, Alon, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant (2019). "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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4149–4158. https://aclanthology.org/N19-1421.

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

Wang, Zijie J., Dongjin Choi, Shenyu Xu, and Diyi Yang (2021). "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*. Online: Association for Computational Linguistics, pp. 47–52. https://aclanthology.org/2021.hcinlp-1.8.

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

Wolf, Thomas, 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 (2020a). "Transformers: State-of-the-Art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Online: Association for Computational Linguistics, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. https://aclanthology.org/2020.emnlp-demos.6.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush (2020b). "Transformers: State-of-the-art Natural Language Processing". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. ACL, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. https://aclanthology.org/2020.emnlp-demos.6.