MISTRA WORKING PAPER

MAPUNGUBWE
INSTITUTE FOR STRATEGIC REFLECTION (MISTRA)

WHY AFRICAN NATURAL LANGUAGE PROCESSING NOW? A VIEW FROM SOUTH AFRICA #AFRICANLP

Reflections on how machines learn to unearth patterns in data and how this capability is then used to try to understand language.

Dr Vukosi Marivate

Vukosi Marivate holds a PhD in Computer Science (Rutgers University) and MSc & BSc in Electrical Engineering (Wits University). He has recently started at the University of Pretoria as the ABSA Chair of Data Science. Vukosi works on developing Machine Learning/Artificial Intelligence methods to extract insights from data. A large part of his work over the last few years has been in the intersection of Machine Learning and Natural Language Processing (due to the abundance of text data and need to extract insights).

02 November 2020

One of the opportunities that Artificial Intelligence (AI) and Machine Learning (ML) (Mitchell, 1997) bring to the emerging Fourth Industrial Revolution (4IR) technologies (Schwab, 2018) is the enhancement of everyday services that we use. AI and ML are behind a large part of how the internet functions today and developments and innovations in these two technologies have brought us numerous services that we now take as the norm. Many of these services we interact with through language (text and speech). The introduction of the search engine heralded a new age in which information on many, sometimes obscure, subjects is accessible at the click of a button (Vise and Malseed, 2005). We have come to expect that these types of services improve over time and bend to our wills. With the 4IR there is an expectation that cyber systems and physical systems will get more intertwined.

One of the ways in which such an expectation is manifesting is the ubiquity of language as an interface with machines. We now speak (literally and figuratively through text) to virtual assistant systems (McLaughlin, 2020), listen to them and make decisions based on their responses or recommendations. We have whole services that let us interact with organisations that are partly automated (we are comfortable interacting with machines through multiple interfaces such as messaging services). How did these technologies come to be? How did machines learn to do what they do? What is the role of local languages in expanding the use of these technologies?

In this chapter, we discuss how machines learn to unearth patterns in data and how this capability is then used to try to understand language. If we are to unlock cyber-physical systems and leave no one behind, then we need to understand and expand on the development of machine-driven local language tools. This is much easier said than done and challenges appear, especially when trying to learn patterns in local languages that do not have much digitised or annotated data to train models to perform natural language tasks.

The aim of this chapter is to motivate why AI/ML technologies for local language are important. The focus will be on South African languages with the aim of encapsulating the challenges local languages face across the African continent in both development and in building ML/AI systems. At the end of the chapter, we try to establish the cause of the challenges local languages face in this area of 4IR technologies and offer suggestions for tackling these challenges. We also highlight work already under way and current successes across the African continent. The chapter uses an adapted Soft Systems (Wheeler et al, 2000) approach to explore challenges and possible solutions in the nexus of machine learning, natural language processing and African languages.

The Soft Systems methodology is used in this chapter to describe the situation we are facing, explain machine learning and natural language processing. The chapter then paints a picture of the challenges that are brought about when considering developing language tools for local languages. Challenges are structured by identifying the factors that lead to having local languages categorised as low resource – that is lacking the data resources required for natural language processing – but at the same time creating a gap in ML/AI systems that can amplify inequity due to biases. Feasible approaches are then proposed to solve these challenges, with examples provided of interventions that fit with these approaches and what they aim to do. First, this chapter explores how machines can understand language.

Machines understanding language: From AI to text

Russell and Norvig (2002: viii) define Artificial Intelligence as 'the designing and building of intelligent agents that receive precepts from the environment and take actions that affect that environment'. Simplified, Artificial Intelligence involves an agent (machine) that resides in an environment. The agent can perceive its environment through sensors and also take actions that change the environment. Creating Artificial Intelligence thus entails building machines that can perceive, act and make decisions in the pursuit of a goal. This goal could be providing guidance on how long it would take to drive to a destination on a rainy day. This requires the perception of the current traffic, the typical commute time and alternative routes. The machine would then have to choose between different routes that require different actions. Making an optimal choice is not trivial (Alpaydin, 2020). What will the machine weigh in making its decision? Fuel economy, time, avoiding tolls? Finally, after the machine makes a choice, a recommendation would be made to the human.

Another example of how machines makes decisions is a general question answering system in which the AI system is given the task of answering the question: 'When did Bafana Bafana (the South African football team) win the African Cup of Nations?'. To answer that question the machine would have to have some form of knowledge database (Brodie and Mylopoulos, 2012). It would have to be able to understand what Bafana Bafana and the African Cup of Nations means, as well as the concept of 'win'. In addition, the machine would need to know that a human would expect the answer the machine gives to involve a date.

If we were to create these machines by programming all of this 'intelligence' from scratch, it would take aeons to come to something that seems very intelligent. We have previously had systems that used pattern matching to exhibit intelligence. An example is a system like ELIZA

(Weizenbaum, 1983), a machine that imitated a therapist, which had been built through simple pattern matching (identifying words a human would type into the computer, and then preparing a 'therapist's' response that matched that word). This way of creating machines that exhibit some form of artificial intelligence does not scale up when we aim for general intelligence. This is where machine learning comes in. Machine Learning (Mitchell, 1997) is the pursuit of getting a machine to learn patterns from data (instead of the patterns being hard coded).

Learning patterns from data provides us with a powerful tool that can lead to impactful solutions to many problems. For example, the use of Machine Learning in automatically identifying fraud on banking transactions means that banks can block accounts immediately when they suspect unusual behaviour (Chandola et al., 2009). The input data is the banking history of the customer and other customers: where they have made a purchases; where they normally purchase; average amounts of purchases, and so on. The goal is to learn which patterns are normal and which are not. That is, can the machine learn what constitutes normal behaviour by the customer and which deviates from the customer's norm?

Another impactful area of application of ML is in medicine. We now have many examples of machine-driven models that can identify specific diseases from medical scans (Thrall et al., 2018), including identifying a tumour that might indicate cancer. The input is the medical scan as an image and the goal is to identify specific markers that may indicate a disease such as cancer.

The examples above are of supervised learning, where the machine gets given data and corresponding labels for each of the data points. The goal of the machine is to then learn connections, specifically the patterns underlying the connections between the input data and the output labels. The representations of data discussed are those of tabular data and images (which are traditionally represented as a set of pixels, each with different numeric values for intensity). But what happens with language? Specifically, how do we deal with text?

Machine learning on text has an application area that is now mostly invisible to all of us: SPAM detection in our email. In SPAM detection the input data is the contents (text, images, links) of our email and other metadata such as who sent the email, and where it was sent from. The label of the email is whether the email is SPAM or 'HAM' – non-spam. The machine takes a representation of this email and then learns from the given labels (when one presses that SPAM button) to distinguish between emails that should be seen by a user or put in the SPAM folder.

SPAM systems have become so good that they are now mostly invisible, while in the earlier days one had to constantly interact with the spam filtering system to prevent spam emails getting through to one's inbox. Now with more data and advances in Machine Learning algorithms, we have powerful SPAM systems that not only look after one's email but are also part of many user-generated content systems online (looking out for hate speech, abuse, illegal content, etc. on the internet). This illustration of a Machine Learning application with text is an example of a natural language processing (NLP) task (Hirschberg and Manning, 2015). This specific illustration is an NLP classification task (SPAM or HAM?).

There are many NLP tasks that we can discuss. Automated text translation is a challenging task that has received much attention recently with many internet services trying to translate different languages. A more challenging task is question answering. This is where a user provides a question to an NLP system, and the system responds with a correct and coherent answer (such as the example provided earlier on Bafana Bafana). Humans can break down the sentence into elements of interest and focus. How do we get a machine to understand the text as well as know what to focus on in its response? How will it even form a response? Where will it get its data to find and form a response? These questions are related to the realm of natural language understanding (NLU), a subset of NLP which focuses on machines/systems comprehending language.

In natural language understanding, a subset of NLP, we want the machine to be able to understand language (comprehension). Through understanding language, we are trying to get the machine to be able to capture intelligence (Goldstein and Papert, 1977). For example, when giving the machine a sample of a paragraph, you would want it to read for understanding and comprehension just like a human. In the earlier days of popular search engines, the queries we entered into the search engine were simple and mostly followed a pattern that the search engine would understand (keyword-based searching). This led to the term Google-Fu to describe the skill to use a search engine (Sem, 2019). Nowadays, search engine queries are as sophisticated as asking a question and getting a response (imagine entering the Bafana Bafana query in 1999 into a search engine). This is a challenging task but one central to the pursuit of having full Artificial Intelligence.

Some of the approaches to NLP that move us closer to general language understanding are language models (Radford, 2019). These language models have followed from the popularity of pre-training of contextual word vectors/models (Pennington et al., 2014) from so-called 'text corpora' or datasets of documents. These approaches take general text (not necessarily connected to the final NLP task) to initialise a model that captures some characteristics of a language. To understand some of the ambitions of language models, imagine giving a machine access to all the English documents available on earth. Through ‘reading’ all of these documents, the machine can learn the structure of the English language (grammar, semantics), comprehend the concepts discussed in different texts and then exploit this ‘knowledge’ in numerous tasks. From this knowledge, machines can then pass through this understanding to a learning pipeline for the final NLP task such as sentiment analysis of statements or clustering (grouping of documents). This is termed transfer learning (Ruder et al., 2019). Transfer learning refers to developing machine learning for one task and using some or all of that model to develop another model for a second, somewhat related, task. Building rich and accurate language models or contextual word vectors requires a large amount of data for pre-training models that can then be used for transfer learning.

As can be seen in Figure 1 and Figure 2, data is needed to be an input into the ML algorithm to produce a trained NLP model. For very complex tasks, we need massive amounts of data. The lack of data becomes a big challenge to solve when we would like to build modern NLP or NLU systems.

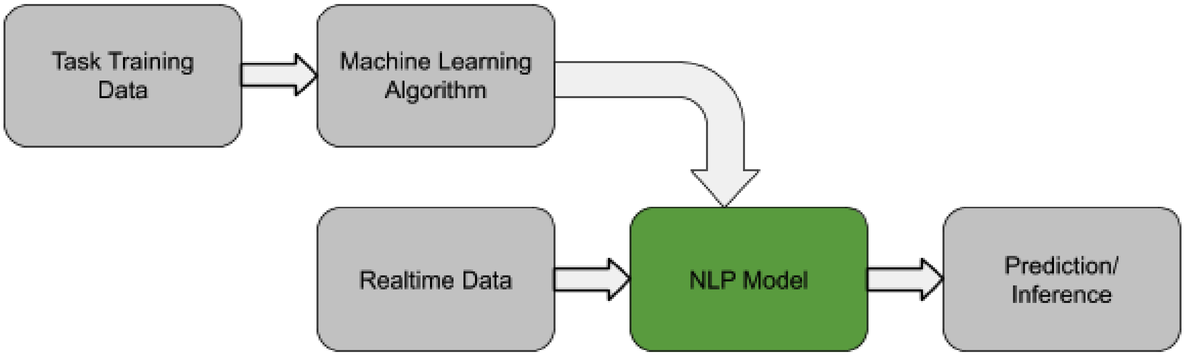

Figure 1. Typical machine learning pipeline (for NLP, training data is some text with or without labels) Source: author

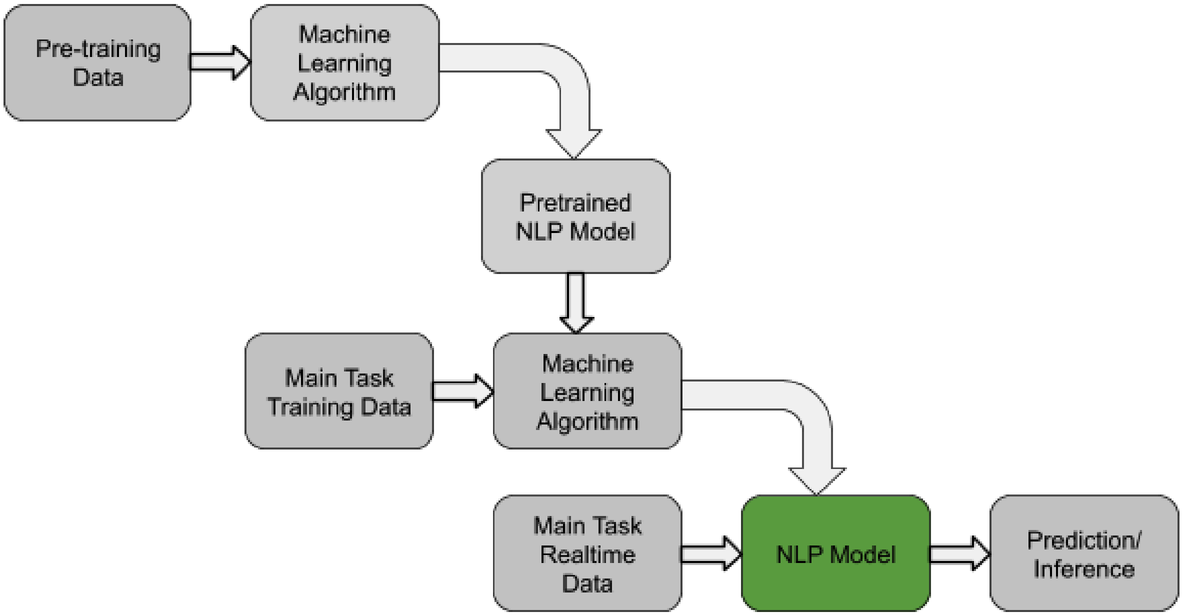

Figure 2. Machine learning pipeline with transfer learning

Source: author

This then brings us to the low representation of African languages, which tend to have small amounts of data available for training these models. For example, if we would like to have a Q&A (question and answer) system that operates in Setswana, isiNdebele or Xitsonga, we need to have a large amount of data that is able to train models and make them robust. This is a challenge that low resource languages have. How then do we talk about innovations in the 4IR in South Africa without talking about language? There can be no such omission. We have to ask, where is the data?

Where is the data?

Machine learning always requires data in one form or another. In NLP we require both more unstructured text data (to pre-train models for reusable representation) and task data to train models for the final task at hand (Figure 2). Systems such as virtual assistants (such as Siri, Amazon Echo, Google Assistant) are mostly available in the world's most resourced languages (Templeton, 2020). These are mostly made up of some European languages, UN languages[i] and medium resourced languages. How can we develop similar systems in South Africa that cover all 11 official languages, or, even better, all the African languages? Africa has the highest language diversity on the planet (Simons and Fennig, 2017). The Niger-Congo language family has 1540 languages, the largest language family in the world (Simons and Fennig, 2017).

Unfortunately, most languages on the African continent are low resourced. Consequently, building NLP systems for these languages is not just a technical challenge, it exposes broader societal challenges in ML systems (Tomašev et al., 2020). We need to be able to source data and also adjust methods to fit these languages (Kann et al., 2019).

The landscape in South Africa

The data available for the nine South African languages (excluding English and Afrikaans) is small for a multitude of reasons. One of these reasons can be seen as inequality. Languages that have enjoyed relative privilege in their development have continued to thrive on the internet even though the internet was heralded as an equalising force. If we look (Table 1) at the relative sizes of local South African languages Wikipedia (Wikimedia Foundation, 2020) sites,[ii] we can get a quick understanding of the uneven distribution of representation of South African languages online (Marivate et al., 2020) as compared to the Statistics South Africa Census.[iii]

Table 1: Wikipedia szes (in terms of number of articles) and corresponding first language speakers in South Africa

Language	Number of Articles	SA first language speakers (%) (2011)
English	6,041,846	9.6
Afrikaans	89686	13.5
Sepedi	8,189	9.1
isiZulu	1,395	22.7
IsiXhosa	1,046	16.0
Setswana	712	8.0
Sesotho	683	7.6
Xitsonga	683	4.5
Swati	504	2.5
Tshivenda	367	2.4
isiNdebele	NA	2.1

Source: Wikimedia Wikipedia Languages; StatsSA 2011 Census

As one can see, for South African languages, the Wikipedia article sizes pale in comparison to English and we dare say Afrikaans. There is simply not enough data to build pre-trained models for each of these languages if Wikipedia, for example, was a source.

Wikipedia is often used to train machine learning and natural language processing systems as it is an easily available and open resource. Wikipedia is also used to create knowledge bases (Lehmann et al., 2015), which are consumed by many services. Knowledge bases are used as ontologies that capture facts and their connections to each other. For example, with a universal knowledge base one is able to extract facts like the president of Mauritius in 1995. The knowledge base will store a concept of a president, the concept of Mauritius as a country, the concept of 1995 as a calendar year and then use these together to answer the query. As such, Wikipedia being unequal has effects far beyond just the language accessibility; it may also be skewing available information to people on the internet (Toyama, 2016). This then requires that we add to our earlier model the impact of societal factors on our NLP model (Figure 3).

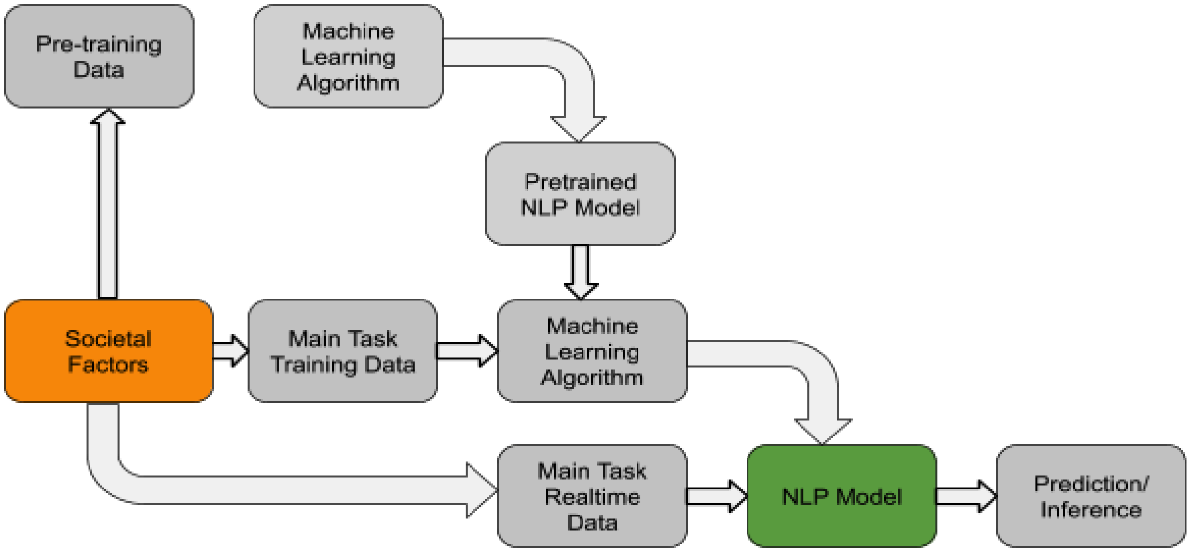

Figure 3. An augmented NLP model, taking into account societal factors acting on data

Source: author

Past experience with South African language processing

Wikipedia is used here as a heuristic to understand a very complex problem. There is much work that has been done in South Africa on building tools for local languages. This section provides an overview of work undertaken thus far in local language processing in South Africa and also covers how recent advances in machine learning are also plotting a new path for local languages.

Work has been done in South Africa on building datasets for Automated Speech Recognition. Work by De Vries et al. (2014) covers collection of 800 hours of speech in all 11 languages. This type of work has also included better understanding of code switching (using multiple languages in the same conversation) for multi-lingual speakers (Modipa et al., 2013) and using soap-opera data speech (Van der Westhuizen and Niesler, 2018). With textual data, work on Setswana corpus creation has been active (Otlogetswe, 2008), focusing on lexicography (building dictionaries). Text to speech has been an active research area by Sefara et al. (2017) and the authors have covered a number of subtopics within their literature. IsiXhosa corpora have been collected for information retrieval (IR) tasks (Packham and Suleman, 2015). There is also work that covers languages such as Xitsonga, isiNdebele, Tshivenda but it is small compared to languages such as isiZulu and isiXhosa.

However, open availability corpora for languages remains a challenge because when one looks at the above papers it is rare to find them on open data repositories for others to use. There are some data repositories that are specifically set up to collect and archive South African language data and these are covered in the next section. The unavailability of open data, open access publications as well as open benchmarks increases impediments to the use and development of tools (Braun and Ong, 2014).

Even so, we are currently going through a renaissance in ML and NLP (Young et al., 2018) that still threatens to leave behind local languages. There has been more focus on African languages over the last decade. One recent example is the panel on African Languages and Digital Humanities: Challenges and Solutions (Petrollino et al., 2019) which discussed the approach of digital humanities to the African languages. This chapter focuses on the use of ML in NLP for African languages.

Recent work (Marivate et al., 2020) focusing on ways to collect and annotate Setswana and Sepedi data shows how one can now investigate the use of pre-trained models (with data collected from various sources) and then create classification models for news. This is ongoing work, with enhancements made to the data using text augmentation methods (which we have shown can work well for short text and low resource scenarios) (Marivate and Sefara, 2020). Results for Setswana and Sepedi news classification are shown in Figure 4. The results shown are on building news classification models that can classify news headline data from SABC radio news stations. Our current focus in this work is to expand the news data from just headlines to full news items as well as to increase the model pre-trained data.

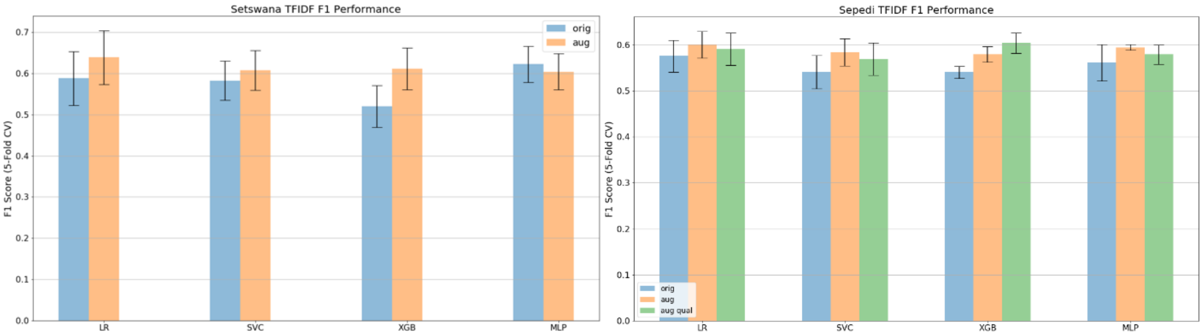

Figure 4: News classification performance (larger is better) for Setswana and Sepedi news headlines from SABC

Source: Adapted from Marivate and Sefara, 2020

Challenges facing local languages

Taking the focus back to the data challenge with local African languages, this chapter now needs to look at why we have such challenges. Martinus and Abbott (2019) discuss some of the problems connected with machine translation for African languages, which also give insight into the challenge with other African language NLP tasks. They highlight the challenges of low availability, discoverability, focus, reproducibility and lack of benchmarks. We will re-examine each of these points and also expand on them.

Low availability

So far this chapter has discussed the challenge of low availability of data in local languages. There are several factors that can be attributed to this. The origins of most of our written language in South African languages was when missionaries went through the country and started documenting (translating) language to convert the local population (Sanneh, 2015). The missionaries largely ignored the scholastic development of the languages they were documenting (De Kadt, 2006: 25–26). This then ties the development of written local languages to the effects of colonialism in South Africa (Tisani, 2005). How then do we get access to digitised versions of local language newspapers, books, and so on, when the inertia caused by the cost of developing scholarship in the languages is a large factor? Initiatives such as the South African Centre for Digital Language Resources (SADiLaR)[iv] are working to use some public funding to move this project forward. SADiLaR focuses on enabling the development of and research into language technologies and language-related studies of South African languages.

Discoverability

For the African language datasets that do exist, how do we find them and get access to them? This is easier said than done. We may have language data resources that are distributed amongst a small set of researchers. We may have language data resources that belong to organisations that are not willing to release these datasets for free. For example, media houses in South Africa have had historical publications in local languages. They might be reluctant to make their data available for machine learning pipelines or language preservation, for commercial reasons.

A question arises: Can we unlock such resources? Having a national broadcaster such as the South African Broadcasting Corporation (SABC) may be one of the vectors to release data. The SABC has news radio stations, and produces news scripts, in all South African languages. This is one of the golden language resources that could be used for not only text NLP but speech as well. Providing a progressive policy for use by researchers and innovators may open up new possibilities. Further, NLP skills are needed inside national broadcasters and traditional media so that they can also add to the research and development of NLP resources and tools. They cannot just be data sources.

Focus

In South Africa, there have been many discussions on the development of indigenous languages (De Kadt, 2006). A major point in these discussions has been the attempt to expand the use of local languages in scholarship and higher education. This is a challenge as society itself is becoming more monolingual (May, 2015). In parallel, we can also look at how African researchers in NLP may be balancing between being able to work on cutting-edge methodology but also wanting to be in the mainstream (Martinus and Abbott, 2019). This then makes it less likely that work on NLP will be focused on indigenous languages as this would be taken as niche and less understood. At the same time, looking at innovations that might be used in the private sector, English would likely solicit more funding (May, 2015). An example of such a situation is how, in South Africa, translation of National Assembly proceedings are translated into English from local languages for inclusion in Hansard which is written in English. This means that when

MPs are talking in a local language it is translated into English only. However, the Hansard is not translated into all the other 10 official languages.

Shifting focus to other government communication, we focus on the publication *Vukuzenzele*. This is a South African government magazine that is published monthly. Even this magazine is supposed to be available in all 11 Languages. In our experience, though, it was only a subset of the English publication (which contains all stories) that are translated. Moreover, some of the English content is republished as is in the local language versions. This is not a critique of the attempt by government communications to provide such a service, but likely indicates the real cost of translation and the consequent choice to focus on the English version. These examples indicate the challenge at hand. Even national newspapers are mostly in English.

Reproducibility and benchmarks

An increasing number of researchers are starting to see the value in making their work available in a reproducible manner. This means making their data available as well as, in the case of NLP/ML, their code. For language tasks, this is important as it not only gives others a benchmark that they can compare against but also a better understanding of the tools that are needed to do work in the area. This challenge is not only in African languages, but in languages in the Global South (Mager et al., 2018). If we aim to grow African NLP, then we need to have more people sharing their work and their data. The international Machine Learning research community is currently at a point where most ML research is shared openly with innovators and researchers who are now able to see the impact of such an approach which is rendering algorithms widely used and easily improved upon.

It is also important to highlight the impact of policy. There is a lot of uncertainty about the use of online services to gather data for training systems on the African continent. The author, being part of the Masakhane NLP project (discussed in depth later), has had to deal with the common questions by researchers and collaborators: 'Can I use data from website X to train my model?'. These questions arise due to researchers not completely understanding copyright and the use of it in ML or NLP tasks. In some states in the EU, there has been a push to have a policy that covers the use of online data for training of text. Countries such as the UK permit such use of data. In South Africa, through NLP research, there is a feeling that doing this increases the likelihood of lawsuits for copyright infringement. This was evident in discussions held through the Masakhane NLP project and questions raised during the first workshop on Resources for African Indigenous Languages (RAIL)[v] in 2020.

How we move forward: Expanding African NLP

We now have an appreciation of the challenges and different facets that make up these challenges. This section discusses several interventions that are needed to improve African NLP and its impact on our societies. This section first discusses how to bring society into the conversation. Next, it discusses ways to improve data collection and curation. Lastly, it surveys the landscape of interventions aimed at expanding skills and practice communities across the continent.

Better public understanding

As we discuss the 4IR, we recognise the need for better public understanding of the underlying emerging technologies. This is not to say every person must have an in-depth understanding of AI/ML/NLP, but it is important that a citizen can understand how these technologies work and how they shape their daily lives. Such understanding will promote more nuanced discourse in public about these technologies and their deployment in society, but, more importantly, it will facilitate a more nuanced discourse around policy. At the moment, without this understanding, asking for clearer policy about the use of online data to train ML/NLP may be seen as unimportant. Without this fundamental understanding, regulations may not change in a timely manner to allow innovations that could completely change some of the technologies for local-use cases. At the same time, researchers may lose opportunities to investigate new tools that may use data that currently would be classified as copyright infringement and IP theft. We do have avenues such as the University of Pretoria (UP) Node of SADiLaR, which works to acquire copyright clearance from publishers before converting local language books to digital formats. Such an approach, though important, will be slower. SADiLaR's approach would also benefit greatly from clearer policy on the use of data for training systems.

Collect, collate and annotate data

The next recommendation is for innovations around data collection, curation and annotation. We need to invest in more data collection, collation and annotation for NLP across the African continent. Such investments would not only serve the aim of increasing the data, but also better preserve the languages. Collection will require content creators to understand ways they can make their content more accessible to machines, not only to humans. One challenge with current typesetting is that it is made for PDFs, which are not an efficient nor accessible way to distribute machine-readable text data. Having even a plain text file of a specific document makes it easier for later consumption. Communicating this sometimes seems counterintuitive as a lot of people who work to create the content are mostly doing it to communicate with the public (humans). But for long-term archiving, having machine-readable text content is not just important, it is essential. This use of machine-readable text is not just for the ML/NLP community, but for archives like libraries as well.

This data then needs to be curated and shared in data archives for future use and expansion. SADiLaR in this case acts as a very good example of such an archive. Many archival data repositories can be used in this manner. Further, as higher education libraries evolve to offer more data archival services across the world, they are expanding their push towards data storage that meets the FAIR standard; NLP can benefit. FAIR data is data with principles of findability, accessibility, interoperability, and reusability (Wilkinson et al., 2016). These principles will go some way towards making sure that the text data that is collected and curated does not suffer some of the challenges that have historically affected low-resource languages.

Finally, we need to annotate the language data for different NLP tasks (Pustejovsky and Stubbs, 2012). This may seem simple but needs careful thought and also has cost implications. To annotate any data, we need to have humans read pieces of text and then make decisions on those annotations. Let’s take for example extracting parts of speech from text. It might be possible to do it programmatically, but we still need humans to validate the work. This has cost implications as data annotators need to be sought and paid. Even in crowdsourcing (obtaining annotations and labels from the public) there is still a cost that we have to factor in and requirements for quality control (Welinder and Perona, 2010). Further, we need to understand what annotation differences might exist for African languages and settings so that we can develop best practice. For example, there might be ideas for recording local folk tales in specific languages from senior speakers of those languages. Doing so might require understanding cultural norms and expectations of what these recordings might mean. This brings us back to the need for an understanding of the technologies so that cultural norms and technological needs can reach some compromise.

The Artificial Intelligence for Development (AI4D)[vi] initiative has recently unearthed a lot of talent and skill with its language data challenge. The challenge called for the collection and curation of African language datasets. The challenge aimed to collect annotated datasets for different African languages and NLP tasks. Through this challenge, researchers could get funding for submitting their collected datasets. The programme revealed some of the challenges language curators have in collecting data. These challenges also mirror some of the challenges discussed before. These included: understanding copyright, digitisation challenges, storage of data and finding innovative ways to convert data. If we solve the data pipeline, we then can move to the NLP task challenges.

Private sector has a role to play in this data collection and African NLP challenge. It collects a multitude of data from users. Let's use the example of social media services which collect large amounts of user-generated content. By amassing this content, the social media services are also collecting local language data. They can improve their services by availing more automated tooling for localised language. One of these tools is services that can better manage abuse and online safety. Abuse here entails concepts such as hate speech, disinformation and predatory behaviour online. For most of the services they have thin (if any) local engineering presence in countries in which they are used. How then do they work to manage automated abuse identification systems? Without more localised knowledge and development their interventions might fall short. Examples of work on NLP tasks on social media in South Africa have mostly focused on English (Featherstone, 2013; Marivate and Moiloa, 2016).

To increase the accessibility of services, more organisations are rolling out interactive services based on text that can take advantage of messaging services becoming ubiquitous. Some of these services offer translated content. Many of the services use automated bots to interact with users. To develop more robust chatbots for local languages, private organisations should assist in creating new datasets as well as benchmarks. We can learn from how many organisations make available their data for different NLP tasks. These then are used to push benchmarks and to create leader boards that ultimately add to the innovation of the field as a whole. An example is the Low Resource Languages for Emergent Incidents project, funded by the US government, to create NLP datasets for some low-resource languages, specifically for disaster management cases (Strassel and Tracey, 2016). Other examples are evaluation competitions such as SemEval (Semantic Evaluation) which have been running since 2001. Every year, SemEval makes available a few semantic evaluation NLP tasks with data and then runs a competition where many groups attempt the tasks and document their work. This then facilitates the creation of a leader board with benchmarks and works to move the field forward. The AI4D programme replicates some of this with a current focus on data collection, but it needs to move to African NLP task challenges that will then spur research and development.

Expanding practice and skill: Building community

To do all of the above, expanding African NLP, we need to be able to muster a lot of resources and people to work together. This will be done by multiple people in diverse ways, but in a coordinated manner. All in all, we need to have an active and growing community to sustain the goals of reaching ideal African NLP. The next section of this chapter focuses on how the rise of AI on the African continent has created fertile soil for the growth of an African NLP community.

In the last decade there has been a rise of initiatives and organisations in Africa that are increasing training and research in the areas of computing in general and artificial intelligence in particular. A number of universities across the continent now have programmes dedicated to Machine Learning, Artificial Intelligence or data science (which tends to have teaching Machine Learning as a core component). We have had the creation of the African Institute of Mathematical Sciences (AIMS), which has boosted the continent-wide availability of mathematical and computing skills at post-graduate level. More recently, AIMS has introduced the African Masters in Machine Intelligence. This traditional-skills pipeline has also incorporated more agile training programmes.

The creation of the Deep Learning Indaba (DLI)[vii] and Data Science Africa (DSA)[viii] initiatives have boosted both the exposure and practice of ML/AI on the continent. The DLI was started with the aim of strengthening African Machine Learning, through Africans being shapers of emerging technologies. DSA aims to create a hub in the network of data science researchers across Africa. These networks are large structures that have worked with each other to provide training opportunities, research collaborations and a community that has grown from strength to strength. Both initiatives also have many NLP enthusiasts who keep working towards different goals but now connect with each other and their networks.

If we look at more regional interventions, we have many examples of programmes. Some are very structured while others are still in their infancy. All in all, they show how young people across the continent are experimenting with the way forward as they try to expand AI/ML practice on the continent. In Nigeria, Data Science Nigeria (DSN)[ix] and AI Saturdays[x] have built a pipeline of students, researchers, engineers and professionals. In South Africa, Explore Data Science internships[xi] and the Data Science for Insight and Decision Enablement initiatives[xii] have been established. In East Africa, Data Science Africa has brought together local enthusiasts and researchers and is working to train the next generation of ML/AI/DS trainers.

After listing some of these interventions, this chapter now looks at how the African NLP community comes together through and with these initiatives. Having identified the talent across the continent, it is important to now move on to innovation and research excellence in what we do. This brings with it more responsibility but also more sustainability: As we show the successes of African NLP, we can garner more support and resources over time. In 2020 there was the AfricaNLP workshop to be held at ICLR 2020 conference. Also, the Resources for African

Indigenous Languages (RAIL) workshop at LREC 2020. These are all indications of the groundswell of interest in the area by practitioners, professionals, engineers and researchers.

One of the obstacles we do have in research across the continent is the need to have large research groups at institutions. Even with many interventions on funding African research, it will take time to build large strong research groups. The Masakahane Machine Translation project (Orife et al., 2020) has provided an alternative template to learn from and improve upon. The Masakhane MT project works on a collaborative distributed research team model. The Masakhane project[xiii] aims to recruit researchers from the African continent to ‘join our effort in building translation models for African languages’. This project combines the challenge of data collection of parallel translation corpora (aligned data in two different languages to allow for a translation task) and the task of training machine translation models. Both parts of the challenge can be taken up, with participants looking or creating new parallel corpora and also fine-tuning machine translation models for benchmarking. The resources to do the research as well as starter material (computational notebooks) are made available freely. The collaborative group meets weekly online and coordinates many of the functions required in the research itself electronically. The evolution and growth of this group will be important as growing pains will highlight some of the areas in which they will need support.

Teaching should also be a priority on the continent. We need to have a pipeline of skills training for cutting edge research and engineering in NLP. Teaching affords us opportunities to try out ideas with students, strengthen teaching networks and also improve upon pedagogy for African NLP. Few higher education institutions across the continent offer natural language processing as a course in their ML/AI/CS/DS curricular. This needs to change. A consortium of teachers has to work together to share their experiences, improve the availability of African language data for their courses and support students in future research endeavours.

Conclusion

This chapter explores the landscape of natural language processing and its connections with artificial intelligence. Artificial intelligence is one of the emerging technologies covered in the Fourth Industrial Revolution. We argue that we cannot completely benefit from AI as an emerging technology in the 4IR without exploring how we improve the state of local language NLP. Local language NLP will lead not only to more inclusive technologies but also to innovations that will drive more AI in South Africa and the African continent. This chapter discussed ways that we can improve local language data creation, collation, curation and annotation to create African NLP task datasets. This requires innovative approaches and partners. We need to tap into resources such as government data and national broadcasters. The private sector also needs to contribute by championing language diversity in their technical systems that use language. To expand the practice of NLP on the African continent, we need to build on the AI/ML community that has been growing over the last few years. This also means training new scientists and engineers as well as expanding research and teaching capacity for NLP. This is an exciting time for African NLP and requires further focus and investment. We need to keep building and pushing for excellence so that our voice can be heard across the world and our languages can be better represented in the evolving and developing technological revolution. We want to live in a world where interacting with intelligent machines in local languages is the norm, not the exception. We look forward to that coming future, a future we all should be working towards and we will realise.

Acknowledgements

We would like to acknowledge the group members and collaborators at the Data Science for Social Impact research group at the University of Pretoria, without whom many of the ideas presented here would not have been explored. We would like to also acknowledge the African machine learning community: Deep Learning Indaba, Data Science Africa, Data Science Nigeria, Masakhane MT, artificial intelligence for Development, Black in AI, and many more. Lastly, we would like to acknowledge Thembekile Marivate for her input in to making the content of this chapter more accessible.

References

Alpaydin, E. 2020 *Introduction to Machine Learning*. MIT Press.

Braun, M.L. and Ong, C.S. 2014. 'Open science in machine learning'. *Implementing Reproducible Research*, 343(2)

Brodie, M.L. and Mylopoulos, J. 2012. *On Knowledge Base Management Systems: Integrating Artificial Intelligence and Database Technologies*. Springer Science & Business Media.

Chandola, V., Banerjee, A. and Kumar, V. 2009. 'Anomaly detection: A survey', *ACM Computing Surveys*, pp. 1–58. doi: 10.1145/1541880.1541882.

De Kadt, J., 2006. 'Language development in South Africa–past and present'. *The Politics of Language in South Africa*. Pretoria: Van Schaik Publishers, 40–56.

De Vries, N.J., Davel, M.H., Badenhorst, J., Basson, W.D., De Wet, et al. 2014. 'A smartphone-based ASR data collection tool for under-resourced languages'. *Speech communication*, 56 (1), 119–131.

Featherstone, C. 2013. ‘Identifying vehicle descriptions in microblogging text with the aim of reducing or predicting crime’, *2013 International Conference on Adaptive Science and Technology*. doi: 10.1109/icastech.2013.6707494.

Goldstein, I. and Papert, S. 1977. 'Artificial Intelligence, Language, and the Study of Knowledge*,†’, *Cognitive Science*, 84–123. doi: 10.1207/s15516709cog0101_5.

Hirschberg, J. and Manning, C. D. 2015. 'Advances in natural language processing’, *Science*, 349(6245), 261–266.

Kann, K., Cho, K. and Bowman, S.R. 2019. 'Towards Realistic Practices In Low-Resource Natural Language Processing: The Development Set’, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. doi: 10.18653/v1/d19-1329.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., et al. 2015. 'DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia’, *Semantic Web*, 167–195. doi: 10.3233/sw-140134

Mager, M., Gutierrez-Vasques, X., Sierra, G. and Meza, I. 2018.’Challenges of language technologies for the indigenous languages of the Americas’, *Proceedings of the 27th International Conference on Computational Linguistics*, 55–69.

Marivate, V. and Moiloa, P. 2016.’Catching crime: Detection of public safety incidents using social media’, *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*. doi: 10.1109/robomech.2016.7813140.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K. and Mokgonyane, T. 2020. ‘Investigating an approach for low resource language dataset creation, curation and classification: Setswana and Sepedi’. *Proceedings of the first workshop on Resources for African Indigenous Languages*, 15–20.

Marivate, V. and Sefara T. 2020. ‘Improving short text classification through global augmentation methods’*. International Cross-Domain Conference for Machine Learning and Knowledge Extraction, https://arxiv.org/pdf/1907.03752.pdf*, accessed 17 July 2020*.*

Martinus, L. and Abbott, J.Z. 2019. ’A Focus on Neural Machine Translation for African Languages’. *arXiv preprint arXiv:1906.05685*.

May, S. 2015.’Contesting Public Monolingualism and Diglossia: Rethinking Political Theory and Language Policy for a Multilingual World’, *Language Policy and Political Theory*, 77–99. doi: 10.1007/978-3-319-15084-0_6.

McLaughlin, M. 2020. ‘How Smart Speakers and Virtual Assistants are Transforming our Lives’. *Lifewire*, https://www.lifewire.com/virtual-assistants-4138533, accessed 13 April 2020.

Mitchell, T.M. 1997 *Machine Learning*. McGraw-Hill.

Modipa, T.I., De Wet, F., and Davel, M.H. 2013. ‘Implications of Sepedi/English code switching for ASR systems’. *Proceedings of the Twenty-Fourth Annual Symposium of the Pattern Recognition Association of South Africa.*

Orife, I., Kreutzer, J., Sibanda, B., Whitenack, D., Siminyu, K. et al. 2020. ‘Masakhane: Machine Translation For Africa‘. *arXiv preprint arXiv:2003.11529*.

Otlogetswe, T.J. 2008. ‘Corpus design for Setswana lexicography’. *Doctoral dissertation, University of Pretoria*.

Packham, S., and Suleman, H. 2015. ‘Crowdsourcing a Text Corpus is not a Game’. *International Conference on Asian Digital Libraries*, 225-234.

Pennington, J., Socher, R. and Manning, C. 2014. 'Glove: Global Vectors for Word Representation’, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543 doi: 10.3115/v1/d14-1162.

Petrollino, S., Nyst, V., Tunde, O., Ngué Um, E., Ekpenyong, M., et al. 2019. ‘African Languages and Digital Humanities: Challenges and Solutions’. *Digital Humanities Conference 2019*

Pustejovsky, J. and Stubbs, A. 2012. ‘Natural Language Annotation for Machine Learning’. *O’Reilly Media, Inc. http://storage.hinterland.nu/webdav/Documents/Data%20Mining/Natural%20Language%20Annotation%20for%20Machine%20Learning.pdf*

Radford, A. Wu, J., Amodei, D., Amodei, D., Jack Clark J., et al. 14 February 2019. ‘Better Language Models and Their Implications’, *OpenAI*, https://openai.com/blog/better-language-models/ , accessed: 11 April 2020.

Ruder, S., Peters, M.E., Swayamdipta, S. and Wolf, T. 2019. 'Transfer Learning in Natural Language Processing’, *Proceedings of the 2019 Conference of the North. Chapter of the Association for Computational Linguistics: Tutorials*, 15-18.doi: 10.18653/v1/n19-5004.

Russell, S and Norvig, P. 2002. *Artificial Intelligence: A Modern Approach (International Edition).* Pearson Prentice-Hall Education International: New Jersey.

Sanneh, L. 2015 *Translating the Message: The Missionary Impact on Culture*. Orbis Books.

Schwab, K. 2018. ‘The Fourth Industrial Revolution (Industry 4.0) a Social Innovation Perspective’, *Tạp chí Nghiên cứu dân tộc*. doi: 10.25073/0866-773x/97.

Sefara, T. J., Manamela, M. J., & Modipa, T. I. 2017. ‘Web-based automatic pronunciation assistant’. *Southern Africa Telecommunication Networks and Applications Conference (SATNAC)*, 112-117.

Sem, S. 16 November 2019. ‘The Beginners Guide to Google-Fu? 10 tricks to be a Google-Fu Blackbelt|, Medium. *Analytics Vidhya*, https://medium.com/analytics-vidhya/https-medium-com-what-is-googlefu-tips-and-tricks-to-be-googlefu-advanced-powersearching-with-google-f7e5661a8bca, accessed 11 April 2020s

Simons, G. F. and Fennig, C. D. 2017. ‘*Ethnologue: Languages of Africa and Europe*. Summer Institute of Linguistics, Academic Publications.

Strassel, S. and Tracey, J. 2016. 'Lorelei Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages’, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 3273–3280.

Templeton, G. 13 May 2020. ‘Language Support in Voice Assistants Compared’. *Globalme.* https://www.globalme.net/blog/language-support-voice-assistants-compared/ accessed 12 April 2020).

Thrall, J. H., Xiang, L., Quanzheng, L., Cinthia, C., Synho D. et al. 2018. 'Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success’, *Journal of the American College of Radiology: JACR*, 15(3 Pt B), 504–508.

Tisani, N. 2005. 'African indigenous knowledge systems (AIKSs): Another challenge for curriculum development in higher education?’, *South African Journal of Higher Education*. doi: 10.4314/sajhe.v18i3.25489.

Tomašev, N. Cornebise, J., Hutter, F., Mohamed, S., Picciariello, A. et al. 2020. 'AI for social good: unlocking the opportunity for positive impact’, *Nature communications*, 11(1), 2468.

Toyama, K. 2016. 'The internet and inequality’, *Communications of the ACM*, 28–30. doi: 10.1145/2892557.

van der Westhuizen, E. and Niesler, T. 2018. ‘A first South African corpus of multilingual code-switched soap opera speech’. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Vise, D. A. and Malseed, M. 2005 *The Google Story*. Random House Digital, Inc.

Weizenbaum, J. 1983. 'ELIZA-a computer program for the study of natural language communication between man and machine’, *Communications of the ACM*, 23–28. doi: 10.1145/357980.357991.

Welinder, P. and Perona, P. 2010. 'Online crowdsourcing: Rating annotators and obtaining cost-effective labels', *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 26(1), doi: 10.1109/cvprw.2010.5543189.

Wheeler, F. P., Checkland, P. and Scholes, J. 2000. 'Soft Systems Methodology in Action: Including a 30-Year Retrospective', *The Journal of the Operational Research Society*, p. 648. doi:10.2307/254201.

Wikimedia Foundation. 2020. '*List of Wikipedias - Meta, Wikimedia Foundation, Inc.* https://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed 12 April 2020.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M. et al. 2016. 'The FAIR Guiding Principles for scientific data management and stewardship', *Scientific data*, 3(1), 1-9

Young, T., Hazarika, D., Poria, S., & Cambria, E. 2018. 'Recent trends in deep learning based natural language processing'. *IEEE Computational Intelligence Magazine*, 55-75.

Endnotes

[i] UN: Official Languages https://www.un.org/en/sections/about-un/official-languages/index.html

[ii] Wikimedia, List of Wikipedias https://meta.wikimedia.org/wiki/List_of_Wikipedias

[iii] Census 2011 Census in brief http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf

[iv] South African Centre for Digital Language Resourceshttps://www.sadilar.org/

[v] The first workshop on Resources for African Indigenous Languages (RAIL) https://www.sadilar.org/index.php/en/news/events/rail2020

[vi] Artificial Intelligence for Development (AI4D) https://ai4d.ai/

[vii] Deep Learning indaba http://deeplearningindaba.com/

[viii] Data Science Africa http://www.datascienceafrica.org/

[ix] Data Science Nigeria https://www.datasciencenigeria.org/

[x] AI Saturdays https://aisaturdayslagos.github.io/

[xi] Explore Data Science https://explore-datascience.net/

[xii] Data Science for Insight and Decision Enablement https://dsideweb.github.io/about/

[xiii] Masakhane: A Focus on Machine Translation for African Languages https://www.masakhane.io/