Masakhane leads the way for low-resourced African languages online

Willem de Vries, Chris Emezue, Bonaventure Dossou

Interviews

2021-07-01

0

0 0 0 0 0 1 1

Though thousands of languages are spoken in Africa, most of Western technology and its abundance of popular applications – which should be widely accessible to all – do not reflect the continent's rich variety of languages. An initiative in which hundreds of volunteers in 20 countries are involved, is currently looking at ways to bring about much needed changes for African languages in technology.

Natural language processing is a discipline within the ever-widening scope of artificial intelligence that draws on many other disciplines to bring human languages into the fold of technological advances, and can do so with far-reaching potential. The prevalence of English in this landscape is limiting, as it does not account for African languages' complexities. To strengthen research into natural language processing in African languages, a grassroots initiative called Masakhane was established in 2019 at the Deep Learning Indaba in Kenya by machine learning engineer Jade Abbott.

More languages for greater accessibility

Masakhane, today a group of more than 400 volunteers from more than 20 African countries, is helping to change predominantly monolingual and homogeneous spaces in technology into something far richer and more accessible for people in and from Africa.

Researchers Bonaventure Dossou and Chris Emezue are part of the Masakhane team, and their work in and approach to African languages in technology have made them recipients of the first Wikimedia Foundation Research Award of the Year.

Award-winning participatory research to change the landscape

Through the award-winning paper, the research writers and the Masakhane community have attempted to change fundamentally how the challenge of low-resourced languages in Africa is approached, following a novel approach for participatory research around machine translation (the translation of texts or speech through technology). This shows how this approach can overcome the challenges these languages face to join the internet and some of the technologies other languages benefit from today, according to the Wikimedia Foundation.

Building speech recognition systems for low-resourced languages

Dossou and Emezue are involved in research into natural language processing, which includes programming that "intuits" the likelihood of a sequence of words in a specific language (called neural machine translation) and speech recognition. Their deep involvement in creating and advancing spaces for African languages includes the creation of the first open-source neural machine translation project between French and Fon, called FFRTranslate; and, for OkwuGbé, they have conducted a comprehensive linguistic analysis of Fon and Igbo, and described the creation of end-to-end, deep neural network-based speech recognition models for both languages. This they see as a step towards building speech recognition systems for low-resourced African languages. Furthermore, they are the cofounders of Lanfrica, a project to connect all African language resources. Dossou and Emezue spoke to Willem de Vries on email about Masakhane and their involvement in natural language processing.

***

How does the scope of the Masakhane project tie in with a broad movement on the continent and elsewhere to reimagine Africa?

Bonaventure Dossou: Africa has always been endowed with so much potential, albeit underestimated by most of the world. Ngũgĩ wa Thiong'o said: "African intellectuals must do for their languages and cultures what all other intellectuals in history have done for theirs." This powerful quote is what I would use to describe the movement going on currently concerning Africa.

There are many areas of growth needed to reach the point where content in African languages is easily accessed on the internet; where one can use Google Assistant, Siri or Alexa freely in an African language; and where African languages are truly represented on the map regarding natural language processing.

Languages reflect current and historical societal infrastructure. In terms of natural language processing, most African languages are classified as "the left behinds", "scraping by" or "hopeful". Some, such as Afrikaans, Kiswahili and Yorùbá, find themselves in the "rising stars" category, according to a recent press release.

There is a lack of natural language processing researchers in Africa. In 2018, only five out of the 2 695 affiliations of participants in the five major natural language processing conferences were from African institutions, according to Masakhane. How do the project's contributors and its participatory research approach engage and strengthen the so-called "low-resourced" languages in Africa?

Chris Emezue: It is important to give an overview first of some of the issues facing the state of natural language processing in less well-known African languages. Among them are a societal lack of focus (the general community, both research and industry, give no focus to these lesser-known African languages because they see no benefit in it), lack of discoverable resources, low creation of public datasets, and little to no reproduction. Masakhane incorporates a participatory approach involving content creators, translators, stakeholders, evaluators and native speakers, among others, for these languages. This helps to foster the inclusion of these languages in technologies.

To what extent does Masakhane collaborate with other similar projects on the continent? What roles do universities have with regard to the Masakhane project? What are the project's areas of interest regarding collaboration and its general aims? Who are the biggest contributors?

Bonaventure Dossou: Masakhane is built on the value of *umuntu ngumuntu ngabantu* (loosely translated from Zulu, meaning "a person is a person through another person" or "I am because you are") and works purely through collaborations among researchers, organisations, institutions of learning and, generally, any interested individuals.

Universities, researchers and companies are free to work with Masakhane. We have had many instances of research collaborations with many organisations in the past few years. The major intersection of all these collaborations (and what one might call the major area of interest) is African languages. In short, the biggest contributors to the Masakhane projects are all Masakhane's members.

How did you get involved in Masakhane? What are your aims within the project?

Bonaventure Dossou: Every single Masakhane member has a motivating story of how they were inspired by Masakhane. The connecting cord in all these stories is the passion the person had for working on African languages.

We (Emezue and Dossou) joined Masakhane when we were undergraduate students in Russia. With our passion for machine learning and natural language processing, it wasn't hard to notice that many research findings and projects were centred mostly on English and other European languages, while there was almost nothing on African languages. We envision a future where language technologies include our native African languages. So, we set out to search for communities dedicated to natural language processing for African languages, and discovered Masakhane.

What are the main challenges Masakhane currently experiences regarding various African languages' low-resourcedness? What has changed since you first got involved with the project?

Chris Emezue: There is a common (mis)conception that low-resourcedness is a data issue, and that the only solution is to get more data. At Masakhane, we have successfully shown that it is a societal issue that requires the whole community to solve it.

Much has changed since we (Dossou and Emezue) joined Masakhane. We have now published 47+ translation models for 35 African languages and (a translation and research tool). We have also given papers at conferences such as EACL, COLING, EMNLP, ACL-WiNLP and ICLR (AfricaNLP), all in 2020 and 2021. To top it all, we won the 2021 Wikimedia Foundation Research Award of the Year for our paper, "Participatory research for low-resourced machine translation: A case study in African languages". It's also important to mention the huge MasakhaNER project (which focuses on named-entity recognition for African languages), as well as the Lacuna grants for four key natural language processing projects.

What developments in the field are you excited about, and why? How does this impact the Masakhane project?

Chris Emezue: Personally, I am excited about the prospects of multilingual machine translation (translation between many languages), as I believe it just may help provide a central solution for translation across the more than 2 000 African languages. I have always envisioned a smooth translation from one African language to another. I am also interested in speech processing. I always opine that the future of African natural language processing lies in its speech, because most of her culture is communicated through an audio medium. At Masakhane, we have a variety of individual and group projects geared towards some of these developments I have talked about.

Bonaventure Dossou: I am excited about everything that natural language processing can bring to African languages: neural machine translation, named-entity recognition, speech (processing, synthesis, recognition and translation) and multilingual translation, among others.
I am also interested in natural language processing applications in healthcare, and I am confident that it will not only help reduce language barriers, but also help the continent get more technological independence. The results of these works, as we have been observing so far, will result in more people being motivated and willing to contribute more to the Masakhane project, which is absolutely great.

What does the roadmap for Masakhane look like?

Bonaventure Dossou: We have machine translation and annotation projects with more than 45 African languages. Our plans include community growth and expanding to more natural language processing tasks beyond machine translation (sentiment analysis and named-entity recognition, to mention a couple). As our goal is to improve reproducibility, we endeavour to create easy-to-use notebooks for our projects so that they can easily be scaled to other African languages. Furthermore, utilising the wide range of native speakers we have, we are making better human evaluation for the natural language processing projects.

How does the usability of natural language processing in English differ to that of any African languages?

Chris Emezue: Many of the existing natural language processing technologies are not able to handle the complexities of African languages, as they were not originally created with African languages in mind.

A good example is diacritisation, which is not prevalent in English, but very impactful in many African languages. Unfortunately, many of the pre-processing methods do not conveniently handle the diacritics, so they simply strip them off.

The good news is that at Masakhane and beyond, researchers are bringing these issues to light and proposing solutions for some African languages (Orife’s Attentive sequence-to-sequence learning for diacritic restoration of Yorùbá language text for Yorùbá is one of many examples). Also, in speech processing, the issues of the tonality of some African languages and code-switching are being discussed and worked on.

We work devotedly on our passion for connecting African languages. Some of the results of this labour are FFRTranslate, OkwuGbé and Lanfrica, which won the 2021 UNESCO & ViVaTech challenge on bridging the barriers in “Cracking the language barrier through data and AI”. All of them tie in with the Masakhane project because we constantly get support and opportunities from all members of the community. Masakhane promotes any and all projects dedicated to African languages.

How would someone in any country in Africa stand to gain from investment in and the successes of Masakhane?

Bonaventure Dossou and Chris Emezue: Masakhane is gaining momentum in the world right now. Research institutions are heavily investing in Masakhane as a central platform for any needed resources to include one or more African languages in their works. We have many research collaborations with Google, Facebook as well as institutions in Africa. Organisations, companies and investors are joining, too, by working together with Masakhane members on several projects which are at the core of African languages (such as including African languages in a medical application or in the agricultural sector). There is much to gain from investing in Masakhane right now, as it is the hottest thing when we talk about natural language processing for African languages.

Buro: MvH

0 0 0 0 0 1 1

0