An overview of digitization of African languages, spellchecking techniques & the progress of spellcheckers globally

Nthabiseng Mashiane
Computer Science Honors
mshnth009@myuct.ac.za

ABSTRACT

In this age where almost everything is digitized, it is important to have the necessary tools to complement efficient digitization of data, in this instance, text data. To this end, it is necessary for word processors to have to be present to ensure the correct documentation of text. This paper gives an overview of the extent to which African languages have been digitized and the process undertaken to digitize respective languages.In addition, this paper briefly touched on the various spellchecking techniques such as minimum edit distance, similarity key as well as n-gram analysis approach. The orthographic rules of a language dictate the approach to use for spellchecking. The paper finalizes by stating the most plausible approach to building a spellchecking tool for isiXhosa.

Keywords

Digitization, spellchecking, orthographic, corpora, lexicon

1. INTRODUCTION

The first spell checker for South African languages was created by D.J Prinsloo in the 1990s. This spellchecker initially worked for isiXhosa, isiZulu, Sesotho sa Leboa and Setswana. Later in 2003, Prinsloo improved the functionality of the spellchecker by increasing the size of the wordlists used for spell checking. Spell Checkers are a ubiquitous tool in this age where almost everything is digitized. They are available for languages which carry commercial value such as English, French, and Spanish etc. as opposed to some indigenous languages particularly in Africa [13].

2. DIGITIZATION OF AFRICAN LANGUAGES

In South Africa, after the birth of democracy, there has been an increase in use of the eleven official languages in official documents as policies have been amended by the state to allow citizens the option of receiving information in their language of choice as opposed to the pre-democratic standard of English/Afrikaans. “Digitization has been defined as the conversion of analogue media to digital form”[1]. It is necessary to digitize African languages as foreign concepts are often imposed on Africa and overwhelm and overpower or heritage [1]. Gibbon et al.[5] further stresses the pertinence of digitization of endangered languages. Furthermore, digitization of African languages allows for the preservation of the heritage and culture and gives emergence of potential areas of research which could possibly increase the number of linguistic experts particularly in South Africa and other African countries. There is a general insufficient digitization of African languages, but there has been an increasing presence in local languages on the web through channels such as blogs and online publishing forums[14]. Bosch et al.[3] note that there are no standards for digitization let alone machine readable lexica which impedes the digitization of these languages.

Bernstein et al.[2] created a web based interactive word processing interface which enables an online community to aid other members with various writing tasks such as editing, proofreading, formatting, etc. called Soylent. In addition, Soylent enables parties to condense text to meet the required word count in the event that the word count is above the limit as well as a proofreading mechanism written using machine learning algorithms. Moving forward, if a South African spellchecker can take this approach for data collection, it could possibly dissolve the issue of the lack of linguistic experts for indigenous languages as a community of native speakers would form online thus, allowing both expansion of the knowledge base and peer review of the data posted. Some issues noted with this approach are that more often than not, reviewers of work submitted can be either those that do the bare minimum or those who go above and beyond their requirements. In both cases, extra work is created for the end user [2] which is undesirable for a spell checking tool.

Machine readable lexica attempt to aggregate all the relevant information of a language in a very structured and compact manner so that the data is re-usable and to ensure that the information abides to some recognized standard [3]. Gibbon et al.[5] studied an endangered African language Ega (largely spoken in the Ivory Coast), in attempt to preserve and digitize it as it was becoming less spoken. To circumvent the lack of standard practice, they devised “better practice” standards for the digitization of the language in attempt to create best practices in documenting, collecting, archiving as well as creating additional tools for training ad discussion platforms. In addition to the urgency of digitization, care must be taken with regards to the digitization process as “many resources become unusable within a decade of their creation” [5] due to the lack of understanding of the language as well as decoding the data from the archived format.

Bantu/African languages are mostly agglutinative. There are models developed for South African languages which were studied by Bell and Bird (2000). This model was unable to accommodate the variety in structure and writing style of the Bantu languages. Bosch et al.[3] , made some modifications to the language model created by Bell and Bird (2000) after having analysed the Bantu language and altered methods of capturing the breakdown of the word (prefix, suffix, affix, infix) in a way that would improve the effectiveness of spell checking mechanisms as well as the possibility of existence of multiple nouns. Bosch et al.[3] also looked at the primary significance of certain words and suffixes as well as distinction of features in nouns and verbs, pertinence/significance reflexive forms and inflections of languages. Throughout the study, approximately 55 indigenous languages were used to evaluate the model which proved the model to work better for some languages as opposed to others. In looking at all of these things, the goal was to "provide and useful and efficient computational resource" [3].

Although texts are now written in and are available in various languages in South Africa, not much of these texts are digitized. This is due to the lack of linguistic experts in South Africa as well as the pool size of native speakers of indigenous languages. In addition, there isn't a standard procedure of data digitization which ensures that the data is captured in a machine readable form [3] which can aid in the creation and advancement of new and existing spell checking tools. Another challenge is the general absence of content generation in the said languages. The tools in existence are proprietary e.g. Spelling Checkers for South African languages, WordPerfect 9, etc. which are often costly and there aren't any effective open source tools in existence and no funding is available for these tools to be created.

3. SPELLCHECKING TECHNIQUES

Spellchecking involves error correction and error detection. Error correction in text has been mainly focused on three areas; non-word error detection, isolated-word error detection and context-dependent word correction [8]. Error detection involves the analysis of pre-generated n-grams from some language corpus, these n-grams may be static or dynamic. Error detection has been successfully implemented while error correction is still progressively being worked on.[8] classifies errors into two; typographical errors which are misspellings and cognitive errors which are errors made by people who don't know how to spell the words, cognitive errors include phonetic errors as well as errors associated with homonyms which can produce a valid word which is erroneous in context. These two types of errors have to be considered when creating a spell checking tool.

Error correction for South African/Bantu languages is still being developed, with efficient algorithms and tools yet to be found. There exist many techniques of error detection, these include minimum edit distance where the algorithm looks for the smallest number of insertions/deletions to correct a word. Another method is the similarity key technique where strings which have a similar spelling are mapped to identical or the same key so that the key of the misspelled word is similar to that of a correctly spelled word or at least gives possible options of the correct word. This seems like a good approach as Damerau (1964) found that 80% of errors in text are a combination of insertion,deletion.substitution and transposition

The structure of African languages is very different from that of the languages catered for on spell checking software (Grover, 2010), and this why not much can be leveraged from the existing spell checkers created for the more commercial languages when creating spell checkers for African languages. String matching algorithms as well as dictionary lookup approaches have been used for spellchecking in Bangla which is looked at in more detail in the next section. None of these methods cater for homonym errors i.e. the errors detected aren't specific to context. UzZaman and Khan [16], were in agreement with Damerau (1964) 's finding that 80% of errors in text are a combination of insertion, deletion, substitution.

4. SPELLCHECKING OUTSIDE OF SOUTH AFRICA

UzZaman and Khan [16] created a Bangla spell checker using generic algorithm which were not tailored to Bangla. In the process, they noted that the orthographic rules are complex and are one of the many reasons why it is difficult to develop good and efficient algorithms to work for the spell checking tool. Bangla, also known as Bengali, is a language spoken in Southern Asia by approximately 210 million people and is one of the top most spoken languages in the world [16].

An Arabic spell checker was created which recognizes common spelling errors and offers suggestions. It was implemented in SICStus Prolog on IBM. The main features of the Arabic language which had to be taken into consideration were computational morphology- which deals with how to derive a new word from an existing one by adding a prefix, suffix or infix and can either change the word category or leave it unchanged. This is referred to as 'morpho graphemic rules' where a word is changed by changing morphological rules. In addition, Arabic has weak and Hamza characters which are characters which are changed by the diacritic of the word. Shaalan et al.[15] summarizes the main five spelling errors in Arabic as follows:

1. Reading errors that occur where an individual is capturing data which is written on paper and misreads some of the data, thus capturing the wrong data. In addition to misreading the data, errors arise from lack of certain characters on the keyboard.

2. Another common error is through transcription where the transcriber hears a different thing from what is being said as there are slight nuances in most of the pronunciation of words which mean very different things. Other reasons for these kinds of errors include the presence of various dialects, the use of slang as well as age.

3. Touch typing errors which would usually be from typists who aren't very experienced. And this would be due to the positioning of the typist on the keyboard.

4. Morphological errors which would arise from a writer who doesn't have much experience.

5. Editing errors which are due to typing errors i.e. insertion, delete, subs.

The Arabic spell checker limited its detection to non-words. After a word is found, various approaches are used to correct the error.

This brief look at what is being done outside of South Africa can give rise to potential hybrid approaches to spellchecking and also give evidence to the extent to which statistical and non-statistical approaches have been successful. Overall, it is important to be aware of the different techniques/approaches used for spellchecking as different approaches tackle different aspects of spellchecking and being aware of these aspects will allow one to create a robust spellchecker.

5. SPELLCHECKING IN SOUTH AFRICA

South Africa has mainly been focused on non-word error detection vs error correction in the existing spell checkers. According to [4], non-word error detection works best when tested against Sesotho sa Leboa vs isiZulu and Afrikaans. Binary n-grams have been used successfully for OCR applications, for spellchecking, probabilistic n-grams are used instead. The higher the order of the n-gram tree, the richer the information [4].

More advanced developments have been looked at which lean towards a more dynamic or automatic error detection where, instead of statically spell checking text after it has been written, spellcheck while they are being typed [7]. Another development is the notion of identifying different languages in texts in order to spellcheck accordingly. The various languages in South Africa usually have diacritics which alter meaning of the word and pose issues when the words are encoded and decoded, more so when the encoding mechanism is not specified and also because there is no current standard of encoding. This issue is usually encountered when the sources of data are varied (Internet, blogs, etc.) as it makes is hard to encode and decode characters.

Jacky Maniacky has developed a tool called Umqageli which means 'diviner' in isiZulu. This tool works for Sesotho sa Leboa, Sesotho, Setswana, Tshivenda. Maniacky (2003) reported that the software does not perform well for the Nguni language group i.e.isiZulu, isiNdebele, isiXhosa and siSwati, neither does it support Xitsonga. Bell and Bird (2000) developed a language model for spell checking South African languages. This model was unable to accommodate the variety in structure and writing style of the Bantu languages but paved a way to the development of a spell checking tool. Bosch et al.(2007), made some modifications to the language model created by Bell and Bird (2000) after having analysed the Bantu language and altered methods of capturing the breakdown of the word (prefix, suffix, affix, infix) in a way that would improve the efficiency of spell checking mechanisms as well as accommodating the possibility of existence of multiple nouns. [3] also looked at the primary significance of certain words and suffixes as well as distinction of features in nouns and verbs, pertinence/significance reflexive forms as well as inflections of languages. Throughout the study, approximately 55 indigenous languages were used to evaluate the model which proved the model to work better for some languages.. In looking at all of these things, the goal was to give useful and efficient computational resources [3].

Ndaba et al.[10] conducted research in the development of spellcheckers where they used a data-driven statistical approach to test the feasibility of using statistical data models in spellchecking. It was concluded that it is indeed a feasible approach to spellchecking. From this, one could speculate that because isiZulu belongs to the Nguni language group, this approach could possibly be mapped onto the other languages in the Nguni group. Ndaba et al.[10] created a spellchecker detector module which uses either word-based n-gram language models or character-based language to analyze a word and check it against an n-gram statistics table for correctness. Upon evaluation, Nadaba et al.[10] found that trigrams gave the best results overall. In addition, although [10] focused on error detection a suggestion to create error corrector module which implemented the minimum edit distance was suggested. Ndaba et al. [10] noticed that the corpora used to test the spellchecker affects the quality of the spellcheck and it is often better to use very recent texts which reflect the orthography of the current language spoken as languages evolve over time. In addition, Jones et al.[6] also points out that variations in a language which can be caused by the evolution of a language (borrowing, neologisms) as well as variations in the representation of a language are some of the main challenges faced when developing a standard isiXhosa spellchecker. To mitigate this, [6] devised set rules to eliminate variation in the development as possible these rules are:

1. Define the language standard of isiXhosa and base the development on that.
2. Have quality control measures in order to validate and clean isiXhosa corpora.

6. CONCLUSION

There are a number of spellcheckers in existence for both English, International and African languages but each have their own shortcomings and challenges. Some of the causes of these shortcomings are the complexity of the language, similarity in phonetics of words, grapheme representations, morphological structure, etc. which make the analysis and modelling very difficult during the spell checking process [16].

This literature review has focused on the digitization of African languages and the general presence of African languages on the internet as well as how digitization affects the creation of the spellchecking tool. This paper also looked as some of the techniques employed in spellchecking tools and the different techniques used by spellcheckers for different languages both in South Africa and outside of South Africa and the importance of the data sources from which the corpora for the respective languages are built.

Overall, we have seen that the statistical approach could possibly be a viable option to create an isiXhosa spellchecking tool [7] and that there can be some preprocessing done to clean the data source prior to building language corpora.

7. REFERENCES

[1] Akinde, T.A., 2007. Digitizing Africa local content: The way forward. *Continental Journal of Information Technology*, *1*, pp.44-50.

[2] Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D.R., Crowell, D. and Panovich, K., 2015. Soylent: a word processor with a crowd inside. Communications of the ACM, 58(8), pp.85-94. DOI=10.1145/2791285

[3] Bosch, S.E., Pretorius, L. and Jones, J., 2007. Towards machine-readable lexicons for South African Bantu languages. *Nordic Journal of African Studies* 16(2): 131–145

[4] De Schryver, G.M. and Prinsloo, D.J., 2004. Spellcheckers for the South African languages, Part 1: The status quo and options for improvement. *South African Journal of African Languages*, *24*(1), pp.57-82.

[5] Gibbon, D., Bow, C., Bird, S. and Hughes, B., 2004. Securing Interpretability: The Case of Ega Language Documentation. In LREC.

[6] Jackie Jones , Kholisa Podile & Martin Puttkammer (2005) Challenges relating to standardization in the development of an isiXhosa spelling checker, South African Journal of African Languages, 25:1, 1-10 DOI= http://dx.doi.org/10.1080/02572117.2005.10587244

[7] Martin, J.H. and Jurafsky, D., 2000. Speech and language processing. *International Edition*, *710*.

[8] Karen Kukich. 1992. Techniques for automatically correcting

words in text. ACM Comput. Surv. 24, 4 (December 1992), 377-439. DOI=http://dx.doi.org/10.1145/146370.146380

[9] Keet, C.M., Khumalo, L. On the verbalization patterns of part-whole relations in isiZulu. *Proceedings of the 9th International Natural Language Generation conference 2016 (INLG'16)*, Edinburgh, Scotland, Sept 2016. ACL, 174-183.

[10] Ndaba, B., Suleman, H., Keet, C.M. and Khumalo, L., 2016, May. The Effects of a Corpus on isiZulu Spellcheckers based on N-grams. In IST-Africa Week Conference, 2016 (pp. 1-10). IEEE. DOI: 10.1109/ISTAFRICA.2016.7530643

[11] Nkomo, D., 2015. Developing a dictionary culture through integrated dictionary pedagogy in the outer texts of South African school dictionaries: the case of Oxford Bilingual School Dictionary: IsiXhosa and English. Lexicography, 2(1), pp.71-99.

[12] Pienaar, W. and Snyman, D.P., 2011. Spelling checker-based language identification for the eleven official South African languages. In Proceedings of the 21st Annual Symposium of Pattern Recognition of SA, Stellenbosch, South Africa (pp. 213-216).

[13] Prinsloo, D.J. and de Schryver, G.M., 2003. Non-word error detection in current South African spellcheckers. *Southern African Linguistics and Applied Language Studies*, *21*(4), pp.307-326.

[14] Scannell, K.P., 2011. Statistical unicodification of African languages. Language resources and evaluation, 45(3), p.375.

[15] Shaalan, K., Allam, A. and Gomah, A., 2003. Towards automatic spell checking for Arabic. In *Proceedings of the 4th Conference on Language Engineering, Egyptian Society of Language Engineering (ELSE), Cairo, Egypt* (pp. 21-22).

[16] UzZaman, N. and Khan, M., 2006. *A comprehensive Bangla spelling checker*. BRAC University.