LT Opportunities and Challenges for Under-Resourced African Languages

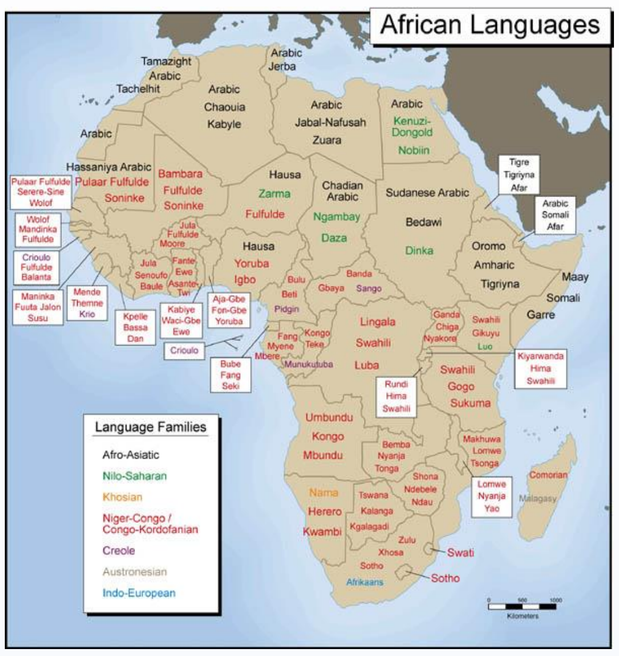
Sunday O. Ojo

Tshwane University of Technology, Pretoria, South Africa

ABSTRACT

Multilingualism and Diversity is the hallmark of African societal heritage, which can be exploited to foster unity in diversity, towards sustainable development. But challenges and opportunities abound in exploiting Language Technology (LT) potentials. Their under-resourcedness and rich contextual semantics features constitute a challenge, given that most of the LTs are validated using richly-resourced languages, such as English as test beds. There are also contextual issues bordering on paucity of internal capacity for exploiting the LT potentials. The poster draws from an on-going experience in working on LTs for some West and Southern African under-resourced languages, eliciting these challenges and opportunities.

THE CONTEXT



- Africa a continent rich in linguistic and cultural diversity
- Over 2000 distinct ethnic groups and associated languages, in 54 nations (Unesco).
- Africa, 14.5% of world population, home to 30.2% of world languages (Simmons and Fennig, 2018).
- Multilingualism and Diversity is an intrinsic socio-cultural heritage of Africa, which should
  - be managed, maintained and exploited,
  - rather than being treated as aberration

    that should be ignored, or treated as an after-thought, in LT development
- Linguistic (Neo-) Colonialism – Anglo-, Franco-phone linguistic dominance;
- Lack of Multilingualism & Diversity (M & D) management capacity, hence so-called official language policies that ignore M & D
- Under-resourced languages-
  - Paucity of lexical semantic resources in quantity and quality for LT development
- Silo uncoordinated language resources development initiatives by individual researchers
- Lack of altruistic motivation , hence commoditization of resources with resistance to open resources

OBJECTIVES

- To elicit challenges and opportunities in African Languages Technology (AfriLT) development
- To proffer a direction in mitigating the challenges and exploiting the opportunities

MATERIALS and METHODS

- Literature-based elicitation of AfriLTs challenges and opportunities
- Experimental AfriLTs development using OpenNLP toolkit
  - Ontology of Yoruba language
  - POS Taggers for Igbo and Setswana Bantu African languages using existing corpora
- Experience- and Literature-based proffering of AfriLTs future direction

RESULTS

Experimental:

- Setswana SVM-based POS Tagger with 96.73% accuracy using governance domain corpus
- Igbo HMM-based POS Tagger with 66.67% accuracy using 13 sentence tokens.
- Striking appropriate balance between AfriLT semantic robustness and computational parsimony, could

  be challenging.

Recommended Future Direction:

- Recognize the need for contextualization of AfriLT provisioning, given:
  - That semantically, every language describes the world in a different way, driven by culture or historical conditions.
  - That Sapir-Whorf Hypothesis holds that the language we speak both affects and reflects our view of the world
  - That state-of-the-art NLP models require large amounts of training data and/or sophisticated language-specific engineering, which is expensive, and requires linguistically trained speakers of the language
  - That African low-resource languages, means they are lacking large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical NLP applications, and so, standard NLP techniques cannot simply be applied, without proper adaptation;
  - That African intellectuals have the cardinal responsibility to render African Languages not only a wider visibility, but also an academic and scientific status through sound Computational Linguistics and NLP research and innovation.

- Given Data Sparcity challenge for AfriLTs the following methodological direction should be explored:
  - Context-driven adaptive use of Supervised Learning methods;
  - Unsupervised feature induction: Brown clustering, Word vectors, etc.
  - Cross-lingual transfer learning – transfer of resources and models from resource-rich source to resource-poor target languages
  - Zero-shot learning – train a model in one domain and assume it generalizes more or less out-of-the-box in a low-resource domain;
  - One-shot learning – train a model in one domain and use only few examples from a low-resource domain to adapt it
  - Joint resource-rich and resource-poor learning using a language-universal representation.
  - Cross-lingual bridging via language lexicons

REFERENCES

Benjamin, M., 2018, May. Hard Numbers: Language Exclusion in Computational Linguistics and Natural Language Processing. In Actes de 11th International Conference on Language Resources and Evaluation (LREC'18), Miyazaki, Japan.

Dibitso, Mary Ambrossine, Pius Adewale Owolawi, and Sunday Olusegun Ojo. "Context-Driven Corpus-Based Model for Automatic Text Segmentation and Part of Speech Tagging in Setswana Using OpenNLP Tool." In International and Interdisciplinary Conference on Modeling and Using Context, pp. 62-73. Springer, Cham, 2019.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig (eds.). 2019. Ethnologue: Languages of the World. Twenty-second edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Eiselen, R. and Puttkammer, M.J., 2014, May. Developing Text Resources for Ten South African Languages. In LREC (pp. 3698-3703).

El-Haj, M., Kruschwitz, U. and Fox, C., 2015. Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. Language Resources and Evaluation, 49(3), pp.549-580.

Pan, S.J. and Yang, Q., 2009. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), pp.1345-1359.

Ruder, S., Peters, M.E., Swayamdipta, S. and Wolf, T., 2019, June. Transfer learning in natural language processing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials (pp. 15-18).

Solorio-Fernández, S., Carrasco-Ochoa, J. A., & Martínez-Trinidad, J. F. (2019). A review of unsupervised feature selection methods. Artificial Intelligence Review. https://doi.org/10.1007/s10462-019-09682-y

Tsvetkov, Y. (2017). Opportunities and Challenges in Working with Low-Resource Languages. Presentation, Language Technologies Institute, Carnegie Mellon University, June 22, 2017

ACKNOWLEDGEMENTS

Mary A. Dibitso, Pius A. Owolawi, and Olamma Iheanetu, with who have been working on AfriLTs under my leadership.