Setswana Word Sense Disambiguation in Machine Translation for Implementation of Pepper Humanoid Robot Instructor

Tebatso Gorgina Moape

Durban University of Technology, Durban, South Africa

ABSRACT

Most of the existing human language technologies are heavily biased toward the natural languages of the developed economies. Ten of which dominate over 80% of the Internet content with the exclusion of African human languages. These technologies cannot be readily applied as humanoid instructors to help communicate education content to students in their mother languages. The focus of this research is on Setswana to English machine translation system with embedded automatic word sense disambiguation (WSD). This research further focuses on resolving the unique challenge of African languages being under-resourced, leading to the digital language exclusion.

INTRODUCTION

- WSD is a crucial component of natural language processing (NLP) tasks with manifold applications. It is an intractably hard open research problem in the disciplines of computational linguistics, computer science, and artificial Intelligence.
- This is because of the contextual nature of ambiguities in human languages and the computational complexity that WSD attracts in NLP.
- The resolution of the WSD problem in the context of an African language such as the proposed Setswana in this study will contribute to the rendition of African languages in wider digital visibility, and academic through sound computational linguistic research.
- In addition, the nature and extent of the morphological richness of a typical African language such as Setswana, resulting in having to grapple with the need for striking an appropriate balance between the robustness and parsimony of the WSD tool, makes this study intellectually challenging due to the nature of the inherent complexity of the WSD task and the lack of language resources in Africa .
- This research aims to make a novel contribution to the theory and practices in human language translation and NLP from the perspective of African languages in general and South African languages in particular.
- This research will be practically implemented in Pepper, a humanoid interactive robot to be used in educational settings to promote learning using African languages and dialects.

SOUTH AFRICAN CONTEXT

		Fully available	Partially available	Non existent
Software: Text	G2P Converter			
	Tokeniser			
	Sentenciser			
	Spelling corrector			
	Full-form normaliser			
	Format normaliser			
	Number normaliser			
	Diacritics normaliser			
	Anonymiser			
	Lemmatiser			
	Stemmer			
	Morphological analyser			
	Morphological synthesiser			
	Part-of-speech tagger/disambiguator			
	Syllabifier			
	Hyphenator			
	Dependency parser			
	Constituent recogniser			
	Chunker			
	Event extractor			
	Named entity recogniser			
	Terminology extractor			
	Topic modelling			
	Sentiment analysis/affect/emotion analyser			
	Referent resolver			
	Word meaning disabmiguator			
	Pragmatic analyser			
	Text generator			
	Summariser			

- South Africa is a country rich in linguistic and cultural diversity.
- Multilingualism and Diversity is an intrinsic socio-cultural heritage of South 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 Language Technology development
- One of the major challenges in language technology development is Under-resourced languages
  - Paucity of lexical semantic resources in quantity and quality for natural language resource development
- Silo uncoordinated language resources development initiatives by individual researchers
- Lack of altruistic motivation , hence ccommoditization of resources with resistance to open resources
- The most recent audit for SA languages indicated that currently, there is no existing sense disambiguator for any of the official SA languages making this research significant to contributing to the language technology landscape.

OBJECTIVES

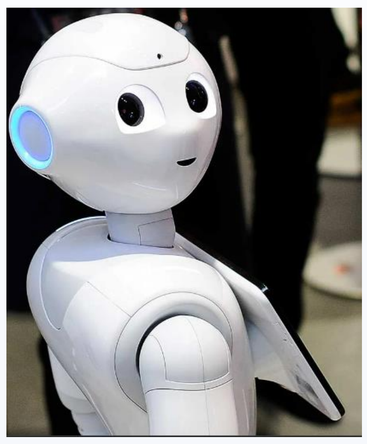
WSD is a crucial component of natural language processing (NLP) tasks with manifold applications. It is an intractably hard open research problem in the disciplines of computational linguistics, computer science, and artificial Intelligence.

- To develop a WSD tool in the context of Setswana-English Machine Translation.
- To practically the developed tool in Pepper, a humanoid interactive robot to be used in educational settings to promote learning using African languages and dialects

METHODS

- Literature Review
  - systematic analysis of the nature and extent of morphology semantic ambiguities in Setswana
- Expert knowledge Sourcing
  - The development of Setswana resource that accommodates morphological richness and ambiguities
- Setswana Language Modelling-Knowledge-based approach using dictionary with Jaccard
  - Develop a language modelling scheme, appropriate for Setswana-English MT WSD
- WSD Tool Development – Python
  - Practically implement and evaluate the language model
- Pepper Implementation – Python
  - Practically implemented in Pepper, a humanoid interactive robot

MATERIALS



- Pepper is a semi-humanoid robot manufactured by SoftBank Robotics, designed with the ability to read emotions.
- It was introduced in a conference on 5 June 2014 and is Python programmable.
- Choregraphe, a multi-platform desktop application, allowing you to: create animations, behaviors and dialogs, test them on a simulated robot, and directly on a real one, monitor and control a robot
- Python, enrich Choregraphe with custom context specific code to improve dialog

EXPETED RESULTS

- Pepper: Humanoid Robot in a python version
  - Pepper, a humanoid interactive robot to be used in educational settings to promote learning using African languages and dialects
- A Setswana Lexical Resource
  - A dictionary that accommodates the morphology, and semantic ambiguities of Setswana
- Setswana – English and English – Setswana disambiguator and machine translation tool
  - A Setswana – English and English – Setswana disambiguator web service machine translation tool

PROGRESS

- Construction of a Setswana word sense knowledge base (Dictionary)
  - A dictionary was developed, a term-to-term dictionary structure was employed as opposed to the conventional word to word and word to definition mapping,
- WSD algorithmic modelling for Setswana
- System Integration: Embedding Setswana WSD algorithmic model into a Setswana-English Machine.
  - Presently evaluating the WSD algorithm and the machine translation system using BLEU evaluation technique.

REFERENCES

Moors, C., Wilken, I., Calteaux, K., & Gumede, T. (2018). Human language technology audit 2018: analysing the development trends in resource availability in all South African languages. Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists, 296–304

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).

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