Current Status, Issues, and Future Directions for Ethiopian Natural Language Processing (NLP)

Seid Muhie Yimam and Chris Biemann

Language Technology Group, Department of Informatics, MIN Faculty, Universität Hamburg, Germany

Current Status of Ethiopian Languages NLP

ACL Anthology Search (phrase search and counts)

Part-of-speech	37500
"Part-of-speech"	25000
Amharic Part-of-speech	293
"Amharic Part-of-speech"	10
Tigrinya Part-of-speech	114
"Tigrinya Part-of-speech"	1
Oromo Part-of-speech	119
"Oromo Part-of-speech"	0
speech recognition	27300
"speech recognition"	11700
Amharic speech recognition	221
"Amharic speech recognition"	9
Tigrinya speech recognition	127
"Tigrinya speech recognition"	6
Oromo speech recognition	90
"Oromo speech recognition"	0

- **Part-of-speech tag works for Amharic**
  - **Amharic:** Getachew (2000) **23 POS tags**, **Mesfin (2001) 25 POS tags**, **Fissaha (2005) 10 POS tags**, Adafre (2005) 10 POS tags, Tachbelie et al. (2009), Gambäck et al. (2009) 30 POS tags, Gebre et al. (2010), Getnet (2015) **more than 100 POS tags**
- **Lemmatizer and Stemmer**: Neshir et al. (2019) removing affixes to a stem, Gasser (2017) apply FST to get lexical/root level

Speech Recognition

- Amharic Speech recognition for translation

  •Abate (2005)
  - 20hr speech data
  - Texts from news paper

  •Woldeyohannis et al. (2017)
  - 20hr Amahric read speech from Tachbelie et al. (2014)
  - 7.43hr read speech corpus using smartphone

  •Gebreegziabher (2019)
  - 90hr speech data
  - Deep learning approaches employed
  - 14.36% word error rate

Machine Translation

- Chala (2010)
  - English-Oromo bilingual translation
  - 20k bilingual and 62k monolingual sentences, BLEU score of 17.74
- Abate et al. (2018)
  - Parallel, bi-lingual English-Ethiopian (7 languages
  - Semitic, Cushtic, and Omotic languages
- Tedla et al. (2016)
  - English-Tigrinya , Phrase based SMT

Word Sense Disambiguation

- Yesuf (2015)
  - Design Amharic WordNet (10,000 synsets and 2000 words)
  - Build Amharic WSD using the WordNet
- Welemichael (2018)
  - Tigrinya Word sense disambiguation
  - Corpus-based approach
- Tilahun (2017)
  - Developed Wordnet (50 polysemous words, 1175 synsets, 1105 words)
  - Word sense disambiguation for Oromo

Information Retrieval and Question Answering

- Question Answering
  - Yimam et al. (2010)
    - Factoid question answering
  - Bete (2013)
    - List based question answering
  - Abedissa et al. (2019)
    - Definition, Biography, and Description QA
- Information retrieval
  - Mequannint (2011)
    - Amharic-English Bilingual Search Engine
  - Getahu et al. (2015)
    - Semantic search (manually built sport ontology)

Text corpus

- A number of online news in Amharic, Tigrinya, and Oromo languages
- More on Social media (Facebook, Twitter, Telegram, WhatsApp…)

Issues

- **Technological issues**

  The Semitic languages are morphologically rich
  - A single Amharic word could give hundreds of morphologically inflicted words
  - How to POS tag such words and how to tokenize texts?
  - There is no proper tokenization tool for Amharic, Tigrinya, Geez, Oromo languages…

  • See this example: Amharic tokenization, it should be more than 3

  ስላለማድረጋቸውስ ፍንጮች ታይተዋል
  [also about them did not perform] [clues] [are observed]
- **Resource-related issues**
  - Papers are usually published in university publication repository
  - Data and code remained fragile, most do not submit their resources, some are submitted to the respective department, some to their supervisor, some at GitHub page
- **Sharing resources**
  - Most researchers do not want sharing resources and codes to others
  - Masakhane [https://masakhane.io/] recently faced a problem to get existing MT parallel datasets

No governing body for Ethiopian NLP

- Research is done based on individual interests
- There are no roadmaps, regular workshops, or conferences particularly targeting NLP for Ethiopian languages
- Data collection and annotation requires a certain budget, there is limited research funding for Ethiopian NLP
- The link between linguistics and computer science departments is not strong, there are no collaborative symposiums or seminars

Other minority languages do not get attention

- **Semitic languages** [1]: Adarigna, **Amharigna**, Argobba, Birale, Gafat, Ge'ez, Guragigna, Chaha group (Chaha, Muher, Ezha, Gumer, Gura), Inor group (Inor, Enner, Endegegna, Gyeto, Mesemes), Silt'e group (Silt'e, Ulbareg, Enneqor, Walane), Soddo group (Soddo, Gogot, Galila), **Tigrigna**, Zay
- **The Cushitic Languages**: Afarigna, Agewigna, Alaba, Arbore, Awngi, Baiso, Burji, Bussa, Daasanech, Gawwada, Gedeo, Hadiyya, Kambatta, Kemant, Konso, Kunfal, Libido, **Oromigna**, Saho, Sidamigna, Somaligna, Tsamai, Werize, Xamtanga
- **The Omotic Languages**: Anfillo, Ari, Bambassi, Basketto, Bench, Boro, Chara, Dime, Dizzi, Dorze, Gamo-Gofa, Ganza, Hammer-Banna, Hozo, Kachama-Ganjule, Kara, Kefa, Kore, Male, Melo, Mocha, Nayi, Oyda, Shakacho, Sheko, Welaytta (Welamo), Yemsa, Zayse-Zergulla
- **The Nilo-Saharan Languages**: Anuak, Berta, Gobato, Gumuz, Komo, Kunama, Kwama, Kwegu, Majang, Me'en, Murle, Mursi, Nera, Nuer, Nyangatom, Opuuo, Shabo, Suri, Uduk

[1] http://www.ethiopiantreasures.co.uk/pages/language.htm

Future Directions

- Start collecting existing resources
  - Currently resources and tools are spread all over the world
- Create data and code repository
  - Create a common repository (Star with existing free repositories such as GitHub)
  - Properly document the resources and their status
  - When resources can not be shared free, arrange possible way of getting access to the resources
- Adapt existing tool
  - Start by integrating and adapting the currently available tools (example integrate the POS tagger to NLTK)
  - Share the code for tools and document the scope of the tools

Prioritized approaches

1. Tokenization and Segmentation
2. Part-of-speech tagging
3. Text and speech corpus
4. Wordnet, thesaurus, and Dictionaries

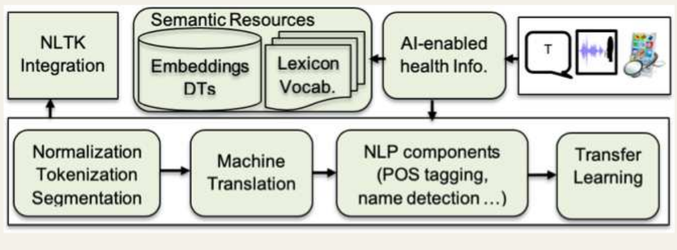
1. Preprocessing and corpus

1. Dependency and constitute parsing
2. Named Entity, Word Sense

2. NLP components

1. Question Answering, Information retrieval, Machine translation, Dialog System
2. Health and Agriculture systems, Legal and financial application

3. NLP Applications



Collaborations, Conferences, workshops

- Establishing research labs in local universities
- Each language should be represented, hence start collecting resources and apply existing models to these languages
- Organize conferences and workshops in a round-robin fashion among the different labs
- Create links with other research organization in Africa