Methods for Amharic Part of Speech Tagging

Björn Gambäck, Fredrik Olsson, Atelach Alemu Argaw, Lars Asker

SICS

Stockholm University

Amharic

- is used for country-wide communication in Ethiopia.
- is spoken by about 30 million people as a first or second language.
- is a Semitic language written from left to right.
- uses a unique script (*fidel*) which has has 33 basic forms and 33 * 7 syllographs
- has a rich verbal morphology based on triconsonantal roots.
- Subject, gender, number, etc., are indicated as bound morphemes.
- nouns (and adjectives) can be inflected for gender, number, definiteness, etc

E.g. sbr : verb forms

	form	pattern
Root	*sbr*	CCC
perfect	*säbbär*	CVCCVC
Imperfect	*säbr*	CVCC
gerund	*säbr*	CVCC
imperative	*sbär*	CCVC
causative	*assäbbär*	as-CVCCVC
passive	*täsäbbär*	täs-CVCCVC

Open Source Taggers

- **TnT**
  - Hidden Markov Model
  - Viterbi algorithm
  - Maximize $P(word_n|tag_n)*P(tag_n|tag_{1...n-1})$
- **SVMTool**
  - Support Vector Machines
  - High dimensional vectors
  - Hyperplane separation algorithm
- **MALLET**
  - Maximum Entropy
  - Linear classifier
  - Log likelihood maximization

Reported performance of the taggers (Wall Street Journal)

Tagger	Performance		
	Overall	Known	Unknown
TnT	96.7%	97.0%	85.5%
SVMTool	96.9 %	97.2 %	83.5 %
Mallet	96.6 %	NA	NA

Experiments and Results

Corpus

Data Set

210k words

1065 Amharic news articles

Tag Set

30 tags – Full tagset by the Ethiopian Languages Research Center (ELRC) [Demeke and Getachew, 2006]

11 tags – Basic tagset by ELRC

10 tags – Alternative tagset by Sisay [Fissaha, 2005]

Tagged Corpus

- Cleaned
- 200,863 words

10 Fold average statistics

Words	Known	Unknown
20,086	17,727	2,359
	88.26%	11.74%

Results

	ELRC	BASIC	SISAY
TnT	85.56	92.55	92.60
STD DEV	0.42	0.31	0.32
KNOWN	**90.00**	**93.95**	**93.99**
UNKNOWN	52.13	82.06	82.20
SVMTool	**88.30**	**92.77**	**92.80**
STD DEV	0.41	0.31	0.37
KNOWN	89.58	93.37	93.34
UNKNOWN	**78.68**	**88.23**	**88.74**
Own folds	88.69	92.97	92.99
STD DEV	0.33	0.17	0.26
MaxEnt	87.87	92.56	92.60
STD DEV	0.49	0.38	0.43
KNOWN	89.44	93.26	93.27
UNKNOWN	76.05	87.29	87.61
Own folds	**90.83**	**94.64**	**94.52**
STD DEV	1.37	1.11	0.69
BASELINE	35.50	58.26	59.61

Discussion

- TnT – Best performance for Known words
- SVMTool – Best performance for unknown words and overall
- MaxEnt – Best performance when it uses own folds

Future Work

- Morphological analysis
- Combining Taggers
- Use external knowledge sources (e.g. machine readable dictionaries)
- Semi supervised / unsupervised learning