NTNU – Trondheim
Norwegian University of
Science and Technology
gamback@idi.ntnu.no

Tagging and Verifying an Amharic News Corpus

Björn Gambäck

Swedish
Institute of
Computer
Science
SICS
gamback@sics.se

Amharic

- Ethiopia's primary language
- About 40 million speakers
- 2nd largest Semitic language
- Unique 275 character script

Amharic Morphology

- Verbs: triconsonantal roots
- N (+ ADJ) inflected for case, gender, number, definiteness
- DEF + CONJ suffixed
- PREP prefixed

E.g. sbr (V, "to break")

	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

Corpus stats

10 folds, averages

WORDS	KNOWN	UNKNOWN
20,086	17,727	2,359
	88.26%	11.74%

Baselines

Most freq.	ELRC	BASIC	SISAY
tag (N)	35.5	58.3	59.6
for word	79.6	83.0	83.1
likely tag	82.6	90.1	90.2

Collocations

(multi-word expressions)

- White-space removal
- One tag for entire MWE

Tachbelie 2010 and Gebre 2010:
- Tag each word separately

(Re-)Tagging and Cleaning the Corpus

Corpus

Corpus Creation
- Walta Information Center news articles
- Web crawl by Stockholm University
- 8715 news items (years 2001-2004) / 1.7M words

Data Set
- 207k words (Unicode and SERA versions)
- 1065 Amharic news articles

Tagging
- Manual tagging by the Ethiopian Languages Research Center (ELRC), Addis Ababa University
- 9 annotators / linguists – typed out by non-linguists

Tag Sets

30 tags – Full tagset by ELRC [Demeke & Getachew, 2006]
11 tags – Basic tagset by ELRC
10 tags – Alternative tagset by Sisay [Fissaha, 2005]

Cleaning

Cleaned Corpus
- Non-tagged parts (e.g., headlines)
- Tagging errors
- Double tags
- Tag misspellings
- Inconsistencies

- 200,863 words
- 33,408 unique wordforms
- 86% of the wordforms have only one possible tag

Amharic Part-of-Speech Tagging

Fissaha 2005:
- CRF; 1k words (test: 10%)
- Reduced 10-class tagset: 70.0% (74.8% with bigrams & MRD)

Tachbelie et al. 2011:
- Comparing TnT, SVMTool, MBT and CRF++; 205k words (test: 5%)
- Best performance overall: 86.3% (SVM)
- Best performance for known words: 87.6% (CRF)
- Best performance for unknown words: 75.1% (SVM)
- Reduced 16-class tagset: SVMTool best (93%)

Gambäck et al. 2009:
- Comparing TnT, SVMTool and MaxEnt (MALLET); 201k words (test: 10%)
- Best performance overall: 88.3% (SVM)
- Best performance for known words: 90.0% (TnT)
- Best performance for unknown words: 78.7% (SVM)
- Best performance if allowed to create its own folds: 90.8% (MaxEnt)
- Reduced 10- and 11-class tagsets: SVMTool best (93%)

Gebre 2010:
- Comparing TnT, Brill and CRF; 207k words (test: 10%)
- Best performance overall: 91.0% (CRF)