A Brief Analysis of Amharic NLP: From POS Tagging to Question Answering

Seid Muhie Yimam

Language Technology Lab

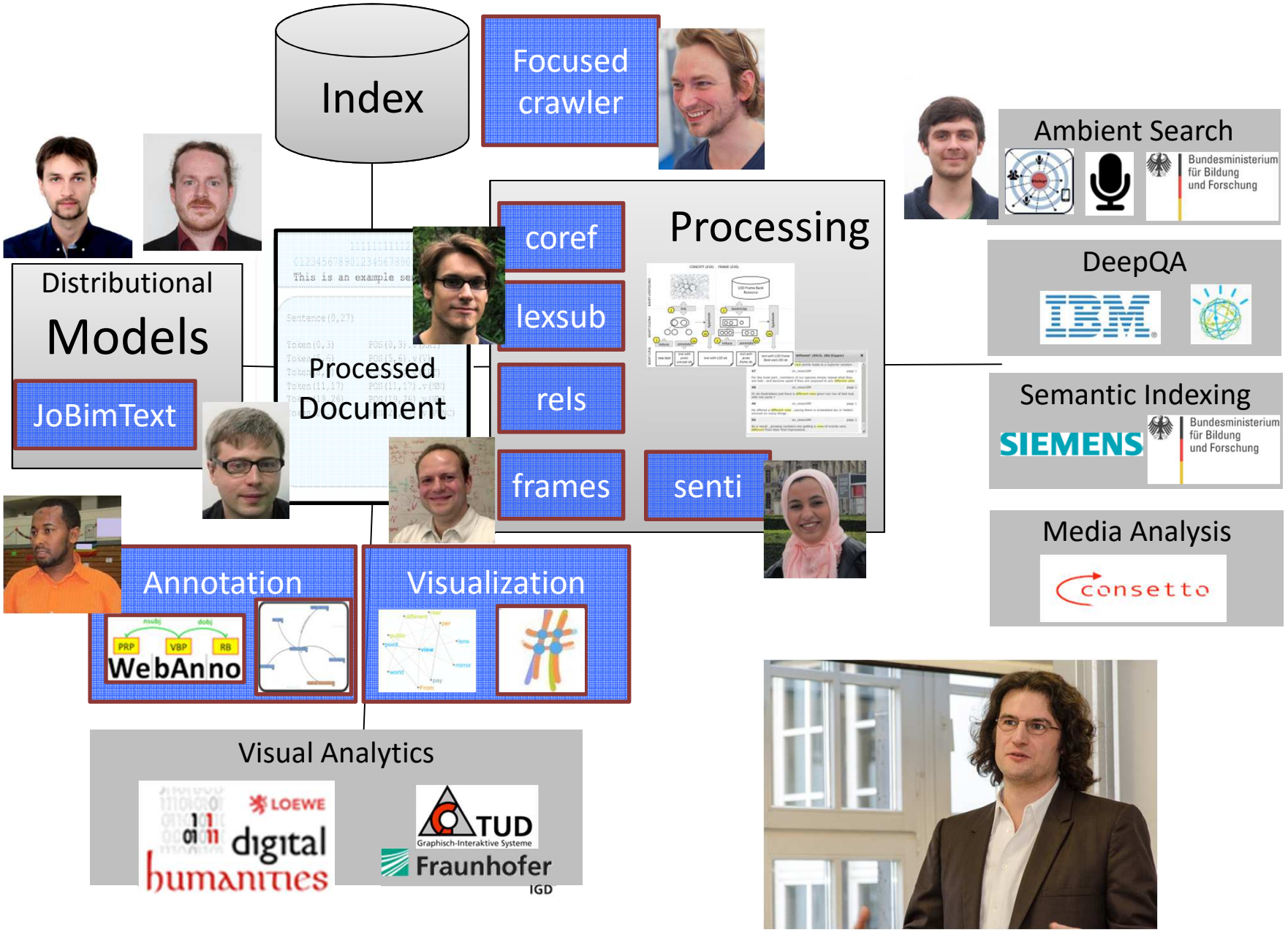
TU Darmstadt

6 September 2016

My profile

- Doctoral researcher at Language Technology lab
- My Topic: Narrowing the loop: integration of **resources** and **linguistic dataset** development with **interactive machine learning**
  - Adaptive annotation in WebAnno
  - Semantic writing aid using contextual paraphrasing

Ecosystem & applications at LT



Disclaimer: I am not a seasoned Amharic NLP researcher. I have done my M.Sc. thesis on Amharic QA in 2008 but rarely participate on Amharic NLP then after. And by no means, this talk is not a detailed analysis of current Amharic NLP.

Outline

- NLP research and applications in Amharic
  - POS tagging, Morphological processing , Spell checking, Named entity recognition, and Questions Answering
- Our Contribution to Amharic NLP
  - POS, QA, Corpora development
- Challenges and bottlenecks in Amharic NLP
- Collaboration with the Ethiopistic department

Introduction

- Amharic writing system: version of the Ge'ez script known as ፊደል (Fidel)
- Ethiopic characters (fidels) have more than 380 Unicode representations (U+1200-U+137F)

	a/ä [a/ɛ]	u [u]	i [i]	a [a]	e [e]	ə [ɨ]	o [o/ɔ]
h [h]	ሀ ha	ሁ hu	ሂ hi	ሃ ha	ሄ he	ህ h(ɨ)	ሆ ho
l [l]	ለ lɛ	ሉ lu	ሊ li	ላ la	ሌ le	ል l(ɨ)	ሎ lo
h/ḥ [h]	ሐ ha	ሑ hu	ሒ hi	ሓ ha	ሔ he	ሕ h(ɨ)	ሖ ho
m [m]	መ mɛ	ሙ mu	ሚ mi	ማ ma	ሜ me	ም m(ɨ)	ሞ mo
s/ś [s]	ሠ sɛ	ሡ su	ሢ si	ሣ sa	ሤ se	ሥ s(ɨ)	ሦ so
r [r]	ረ rɛ	ሩ ru	ሪ ri	ራ ra	ሬ re	ር r(ɨ)	ሮ ro
s [s]	ሰ sɛ	ሱ su	ሲ si	ሳ sa	ሴ se	ስ s(ɨ)	ሶ so
sh/š [ʃ]	ሸ ʃɛ	ሹ ʃu	ሺ ʃi	ሻ ʃa	ሼ ʃe	ሽ ʃ(ɨ)	ሾ ʃo
k'/q [k']	ቀ k'ɛ	ቁ k'u	ቂ k'i	ቃ k'a	ቄ k'e	ቅ k'(ɨ)	ቆ k'o

	a/ä [a/ɛ]	u [u]	i [i]	a [a]	e [e/ɛ]	(ə) [ɨ/ʉ]	o [o/ɔ]
h/ḫ [h]	ኸ hɛ	ኹ hu	ኺ hi	ኻ ha	ኼ he	ኽ h(ɨ)	ኾ ho
w [w]	ወ wɛ	ዉ wu	ዊ wi	ዋ wa	ዌ we	ው w(ʉ)	ዎ wo
ʻ/ʿ [ʔ]	ዐ ʔa	ዑ ʔu	ዒ ʔi	ዓ ʔa	ዔ ʔe	ዕ ʔɨ	ዖ ʔo
z [z]	ዘ zɛ	ዙ zu	ዚ zi	ዛ za	ዜ ze	ዝ z(ɨ)	ዞ zo
zh/ž [ʒ]	ዠ ʒɛ	ዡ ʒu	ዢ ʒi	ዣ ʒa	ዤ ʒe	ዥ ʒ(ɨ)	ዦ ʒo
y [j]	የ jɛ	ዩ ju	ዪ ji	ያ ja	ዬ je	ይ j(ɨ)	ዮ jo
d [d]	ደ dɛ	ዱ du	ዲ di	ዳ da	ዴ de	ድ d(ɨ)	ዶ do
j/ǧ [ʤ]	ጀ ʤɛ	ጁ ʤu	ጂ ʤi	ጃ ʤa	ጄ ʤe	ጅ ʤ(ɨ)	ጆ ʤo
g [g]	ገ gɛ	ጉ gu	ጊ gi	ጋ ga	ጌ ge	ግ g(ɨ)	ጎ go

qh	ቐ	ቑ	ቒ	ቓ	ቔ	ቕ	ቖ
[ʁ']	ʁ'ɛ	ʁ'u	ʁ'i	ʁ'a	ʁ'e	ʁ'(ɨ)	ʁ'o
b	በ	ቡ	ቢ	ባ	ቤ	ብ	ቦ
[b]	bɛ	bu	bi	ba	be	b(ɨ)	bo
t	ተ	ቱ	ቲ	ታ	ቴ	ት	ቶ
[t]	tɛ	tu	ti	ta	te	t(ɨ)	to
ch/č	ቸ	ቹ	ቺ	ቻ	ቼ	ች	ቾ
[ʧ]	ʧɛ	ʧu	ʧi	ʧa	ʧe	ʧ(ɨ)	ʧo
h/ḫ	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
[h]	ha	hu	hi	ha	he	h(ɨ)	ho
n	ነ	ኑ	ኒ	ና	ኔ	ን	ኖ
[n]	nɛ	nu	ni	na	ne	n(ɨ)	no
ny/ñ	ኘ	ኙ	ኚ	ኛ	ኜ	ኝ	ኞ
[ɲ]	ɲɛ	ɲu	ɲi	ɲa	ɲe	ɲ(ɨ)	ɲo
ʾ/ʔ	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
[ʔ]	ʔa	ʔu	ʔi	ʔa	ʔe	ʔɨ	ʔo
k	ከ	ኩ	ኪ	ካ	ኬ	ክ	ኮ
[k]	kɛ	ku	ki	ka	ke	k(ɨ)	ko

t'/ṭ	ጠ	ጡ	ጢ	ጣ	ጤ	ጥ	ጦ
[t']	t'ɛ	t'u	t'i	t'a	t'e	t'(ɨ)	t'o
ch'/č̣	ጨ	ጩ	ጪ	ጫ	ጬ	ጭ	ጮ
[ʧ']	ʧ'ɛ	ʧ'u	ʧ'i	ʧ'a	ʧ'e	ʧ'(ɨ)	ʧ'o
p'/p̣	ጰ	ጱ	ጲ	ጳ	ጴ	ጵ	ጶ
[p']	p'ɛ	p'u	p'i	p'a	p'e	p'(ɨ)	p'o
ts'/ṣ	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
[ts]	ts'ɛ	ts'u	ts'i	ts'a	ts'e	ts'(ɨ)	ts'o
ts'/ṣ́	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ
[ts]	ts'ɛ	ts'u	ts'i	ts'a	ts'e	ts'(ɨ)	ts'o
f	ፈ	ፉ	ፊ	ፋ	ፌ	ፍ	ፎ
[f]	fɛ	fu	fi	fa	fe	f(ɨ)	fo
p	ፐ	ፑ	ፒ	ፓ	ፔ	ፕ	ፖ
[p]	pɛ	pu	pi	pa	pe	p(ɨ)	po
v	ቨ	ቩ	ቪ	ቫ	ቬ	ቭ	ቮ
[v]	vɛ	vu	vi	va	ve	v(ɨ)	vo

ቊ k'ʷi	ቋ k'ʷa	ቌ k'ʷe	ቍ k'ʷɛ	ኊ hʷi	ኋ hʷa	ኌ hʷe	ኍ hʷɛ	ኲ kʷi	ኳ kʷa	ኴ kʷe	ኵ kʷɛ	ጒ gʷi
ጓ gʷa	ጔ gʷe	ጕ gʷɛ	ሏ lʷa	ቧ bʷa	ዟ zʷa	ጧ t'ʷa	ሟ mʷa	ቷ tʷa	ዧ ʒʷa	ጯ ʧ'ʷa	ሯ rʷa	ቿ ʧʷa
ጇ ʤʷa	ጿ ts'ʷa	ሷ sʷa	ኗ nʷa	ዷ dʷa	ፏ fʷa	ሿ ʃʷa	ኟ ɲʷa	ፘ rʲa	ፙ mʲa	ፚ fʲa	ኧ ʔɛ	

፣	።	፥	፤	፦	፧
comma	full stop / period	colon	semi-colon	preface colon	question mark (no longer used)

፩	፪	፫	፬	፭	፮	፯	፰	፱	፲
አንዶ	ሁለት	ሶስት	አራት	አምስት	ስዶዶስት	ሰባት	ስምንት	ዘጠኝ	አስር
and	hulätt	sost	aratt	ammǝst	sǝddǝst	säbatt	sǝmmǝnt	zäṭäňň	assǝr
1	2	3	4	5	6	7	8	9	10
፳	፴	፵	፶	፷	፸	፹	፺	፻	፲፻
ሀያ	ሰላሳ	አርባ	ሀምሳ	ስልሳ	ሰባ	ሰማንያ	ዘጠና	መቶ	ሺ
haya	sälasa	arba	hamsa	sǝlsa	säba	sämaňa	zäṭena	mäto	ši
20	30	40	50	60	70	80	90	100	1000

Amharic morphology [1/2]

- Amharic is morphologically complex

Example Noun: መሬት (meret) ~ land

መሬቶች (meretoch) ~ lands

መሬቶቻችን (meretochachin)~ our lands

ስለመሬቶቻችን (silemeretochachin)~ about our lands

ስለመሬቶቻችንስ? (silemeretochachins)~ what about our lands?

Amharic morphology [2/2]

Example Verb

መስበር (mesber) ~ break

ሰበራችሁ (seberachihu)~ you break (s/t)

ተሰበራችሁ (teseberachihu) ~ you are broken

ተሰባበራችሁ (tesebaberachihu) ~ you are broken (to portions)

ስላልተሰባበራችሁ (silaltesebaberachihu) ~ as you are not broken

ስላልተሰባበራችሁም (silaltesebaberachihum) ~ and as you are not broken

A single verb may consist time (past, present, and future), gender (male and female), action (command, statement, invitation) and negation (not).

HORNMORPHO 2.5[1]

- morphological analysis and generation of Amharic and Oromo verbs and nouns and Tigrinya verbs

```
>>> l3.anal('am', 'ስላልተሰባበራችሁም')
Loading morphological data for Amharic ...

word: ስላልተሰባበራችሁም
POS: verb, root: <sbr>, citation: ተሰባበረ
 subject: 3, sing, masc
 object: 2, plur
 grammar: perfective, iterative, passive, relative, definite, negative
 preposition: sIle, conjunctive suffix: m
POS: verb, root: <sbr>, citation: ተሰባበረ
 subject: 2, plur
 grammar: perfective, iterative, passive, relative, negative
 preposition: sIle, conjunctive suffix: m
```

[1] http://homes.soic.indiana.edu/gasser/Research/software.html

Spell checking

- No notion of “spelling” in Amharic [1]
  - “if a word sounds right when read aloud then it was rightly written”
  - “ውኃ” vs “ዉሀ”, “ታህሳስ” vs “ታኅሣሥ” -> acceptable
  - “ዓዲሥ ዐበባ ዒትዮጵያ” vs “አዲስ አበባ ኢትዮጵያ not acceptable

[1] https://arxiv.org/ftp/cs/papers/0408/0408052.pdf

Spell checking

- Different levels (elementary, intermediate, advanced)
- “ባል”, “መልክት”, “እንቁጣጣሽ”, “አይን”, “አሳ”, “ያገር”
- “በዓል”, “መልዕክት”, “እንቍጣጣሽ”, “ዓይን”, አሣ”፣“የአገር”
- “በዐል”, “መልእክት”, “ዕንቍጣጣሽ”, “ዐይን”, “ዐሣ”, “የሀገር”

Spell checking

- Metaphone algorithm and an edit distance algorithm [1]

1. Get *WORD* in the dictionary
2. Compute Amharic Metaphone code for *WORD*, *W_CODE*
3. IF *WORD* has no similar *W_CODE* with a previously hashed word:
   Add *W_CODE* as a key and *WORD* as value in a hash table, *H_DICT*
   ELSE:
   Add *WORD* as the next value in the respective key
4. IF the dictionary has another *WORD*:
   GOTO 1

[1] http://thirdworld.nl/development-of-an-amharic-spelling-corrector-for-tolerant-retrieval

Results

Rank of Correct Suggestions	Count	Percentage
Top One	55	67.1%
Top Three	61	74.4%
Top Five	67	81.7%
N/A	15	18.3%
Total	**82**	**100%**

POS tagging [1/3]

- Getachew (2001) : - Hidden markov models
  - definition of a tagset of 25
- Adafre (2005): stochastic model based on conditional random fields
  - revised Getachew's tagset and reduced it to 10
  - average accuracy of 74%
  - used dictionaries of affixes
  - some 15,000 entries with their POS tags (Noun, Verb, Adjectives, Adverb, and Adposition)
- (Demeke and Getachew, 2006):
  - corpus of 1065 news articles
  - 31 parts of speech
  - Multi-word expressions use separate tag

POS Tagging [2/3]

- Gamback et al. (2009): using TnT, SVMTool, and Mallet
  - Accuracy of 85.56% for TnT, 88.30% for SVM and 87.87% for MaxEnt
- Tachbelie and Menzel (2009): using TnT and SVMTool models
  - accuracies of 82.99% for TnT and 84.44% for SVM

POS Tagging

- Binyam (2011)
  - clean the WIC data
  - Use vowel patterns and the radicals as POS features
  - State-of-the-art tagging models (CRF++, LIBSVM, and TnT)

POS tagging problems

- The corpus used is usually small in size
- The quality of the corpus is poor
  - A lot of inconsistency
- Due to the agglutinative nature of the language
  - Fine-grained tagsets – difficult to annotate large data set
  - Coarse-grained tagsets – more ambiguous and might be wrong
- Lack of annotation guidelines
  - Computer programmers do not know much of the linguistic theory
  - Linguists do not know much of the programming

Definitive QA (Teshome 2013) [1]

- Based on surface text pattern method
  - Design patterns to discover a set of definition-related text patterns from the Amharic legal corpus.
  - Extract a collection of concept-description pairs from a target document
  - Apply the definition extraction to return answer to a given question.
- Achieved F-measure of **78.8**%

[1] http://etd.aau.edu.et/bitstream/123456789/8300/1/Wondwossen%20Teshome%202013.pdf

QA for list questions (Bete 2013) [1]

- Closed domain (Ethiopian tourism)
- Hypothesis:
  - answers to a list questions have same semantic entity class
  - Answers that co-occur within the sentences of the documents are related to the target
  - The question and sentences containing the answers share similar context.

[1] http://etd.aau.edu.et/bitstream/123456789/8587/1/Brook%20Eshetu%20Bete.pdf

Amharic Named Entity recognition [1]

- Machine learning approaches
  - Conditional random fields
  - 80.66% of F-measure achieved
- Amahric NE
  - no case information like English
  - Clue words for each NE classes are used
    - ፕሬዚዳንት President, አቶ Mr. , ወ/ሮ Mrs
    - ካፒታል ሆቴል Capital Hotel, ሐረማያ ዩኒቨርሲቲ Haramaya University ፕሬዚዳንት ቢሮ, Office of president , የተባበሩት መንግስታት ድርጅት, United Nations Organization

[1] http://etd.aau.edu.et/bitstream/123456789/8579/1/Besufikad%20Alemu%20final.pdf

NER challenges for Amharic

- Ambiguity:
  - ፀሐይ (Sun) person name as well the object sun
  - Some names are used as both person and location names
- Spelling variations
  - “ፀሐይ”, “ጸሐይ”, “ፀሀይ”, “ጸሀይ” ~ sun
- Lack of capitalization

Our contribution

Factoid QA system for Amharic

- The first QA system for Amharic
- Components
  - Document pre-processing
  - Question processing
  - Document retrieval
  - Sentence/paragraph re-ranking
  - Answer selection modules.

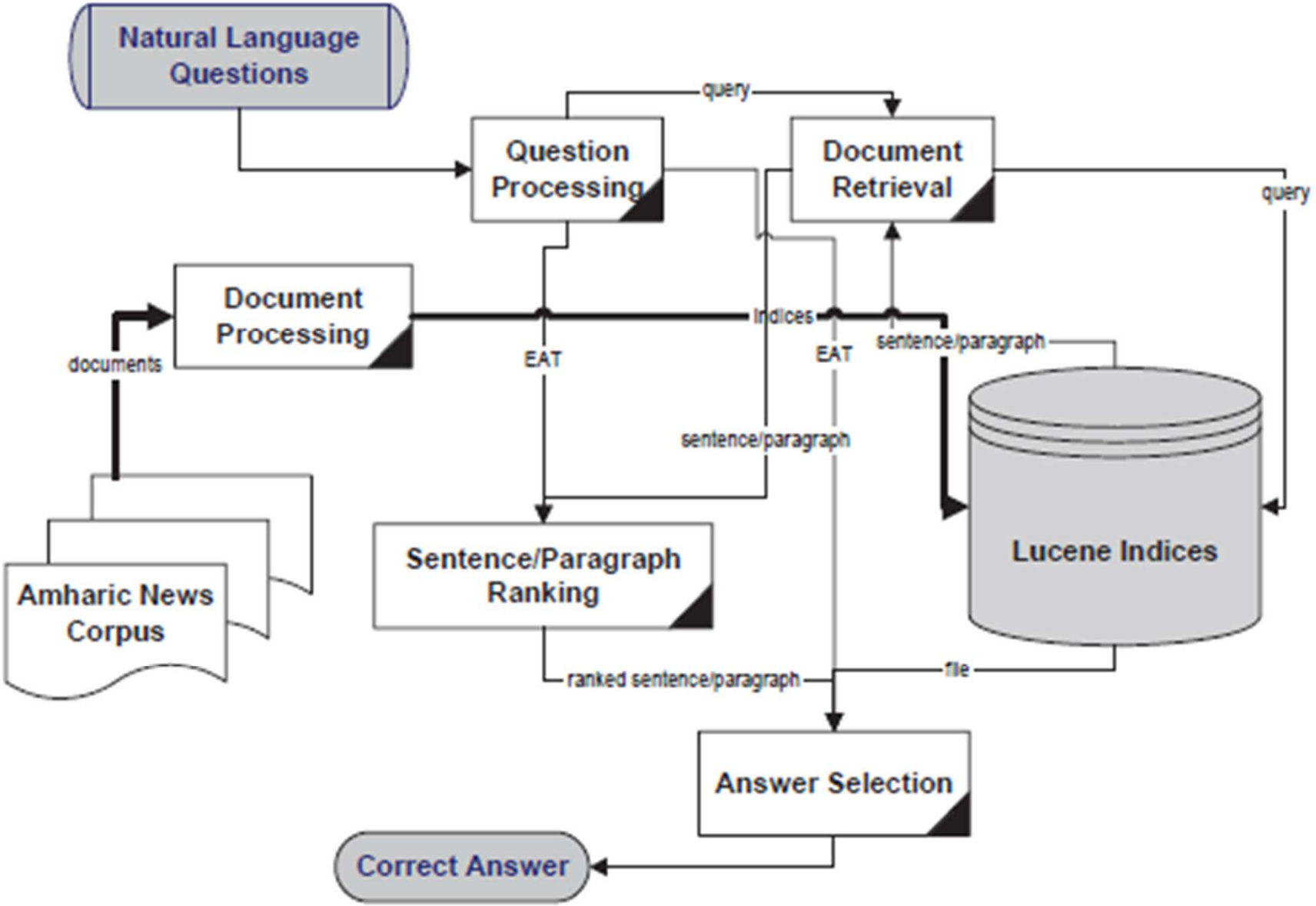
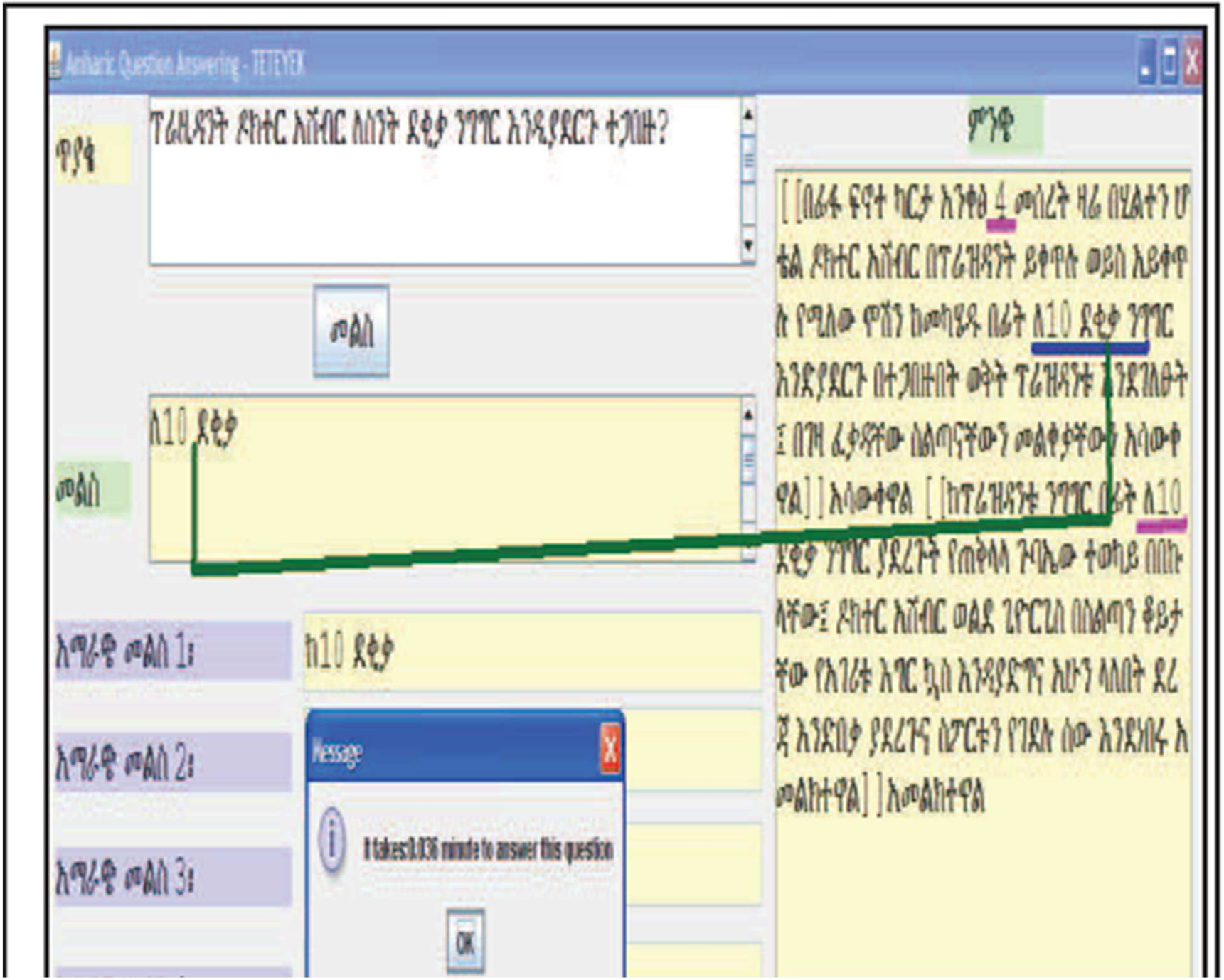


Figure 1: Architecture of the system


Amharic Question Answering - TETEYEN
ጥያቄ
ፕሬዚዳንት ዶክተር አሸብር ለሰንት ደቂቃ ንግግር እንዲያደርጉ ተጋበዙ?
መልስ
መልስ
ለ10 ደቂቃ
ምንጭ
[[በፊፋ ፍፃሜ ካርታ እንቀፅ ፯ መሰረት ዛሬ በሂልተን ሆቴል ዶክተር አሸብር በፕሬዝዳንት ይቀጥሉ መይስ እይቀጥሉ የሚለው ጥሸን ከመጣፄዱ በፊት ለ10 ደቂቃ ንግግር እንደያደርጉ በተጋበዙበት ወቅት ፕሬዝዳንቱ እንደገለፁት ፤ በገዝ ፈቃዳቸው ለልጣናቸውን መልቀቃቸውን አሳውቀዋል]] አሳውቀዋል [[ከፕሬዝዳንቱ ንግግር በፊት ለ10 ደቂቃ ንግግር ያደረጉት የጠቅላላ ጉባኤው ተወካይ በበኩላቸው፤ ዶክተር አሸብር መልደ ገቦርገስ በስልጣን ቆይታቸው የአገሪቱ እግር ኳስ እንዲያድግ አሁን ላለበት ደረጃ እንደበቃ ያደረጉና ሀገርቱን የገደሉ ሰው እንደነበሩ አመልክተዋል]] አመልክተዋል
አማራጭ መልስ 1፡
ከ10 ደቂቃ
አማራጭ መልስ 2፡
አማራጭ መልስ 3፡
Message
It takes:0.026 minute to answer this question
OK

Results

- **12000** question sets have been collected from the Web, Ethiopian television games and from questionnaire respondents.
- Question classification **89%**
- Retrieval
  - Document based **97%** recall
  - Sentence based **93 %** recall
- Answer selection **72%** recall

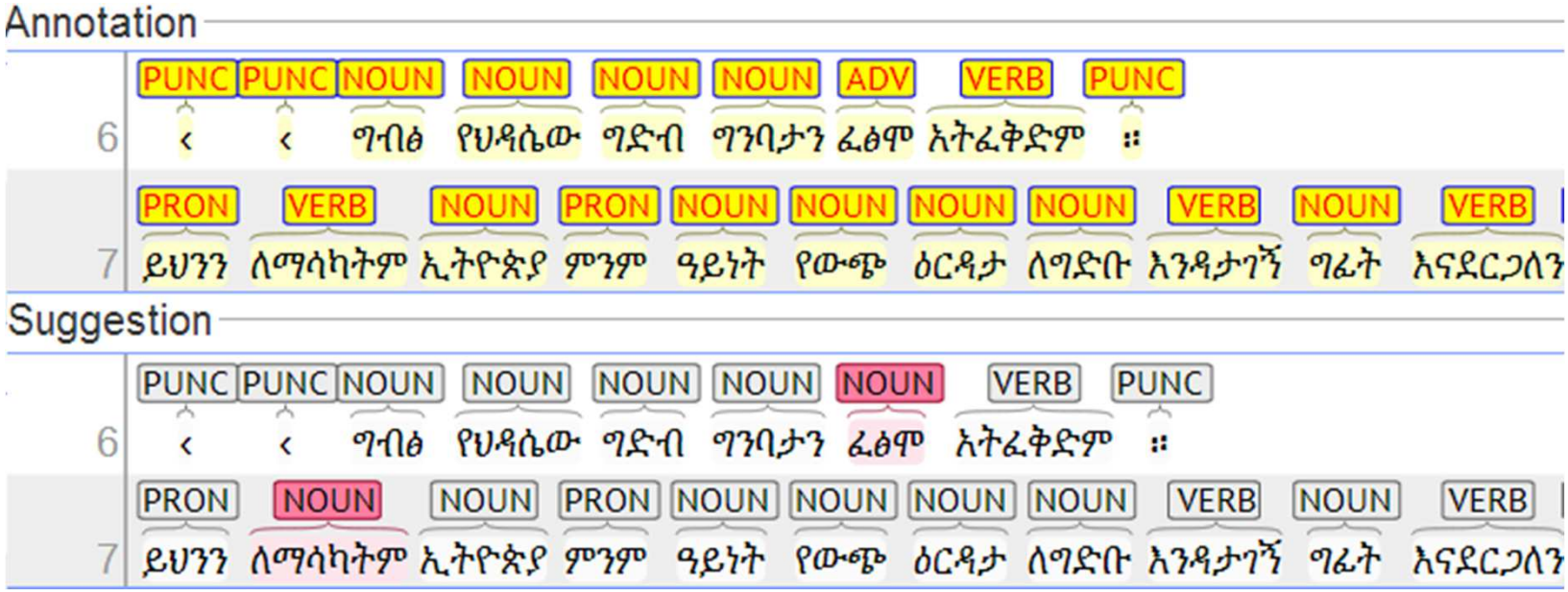
POS Tagging

- POS Automation [1]
  - Using WebAnno automation and correction module
  - manually annotate Amharic documents
  - Designed 11 POS tags equivalent to the universal POS tags

[1] http://www.aclweb.org/anthology/P/P14/P14-5016.pdf

POS tagging automation

- Initially manually annotated 21 sentences
- Iterative automatic annotation suggestion until 300 sentences are annotated.
- We obtained an F-score of 0.89 with the final model.



Unsupervised POS tagging (Yemisrach - 2015)

- Co-advising a student master thesis work at Gondar University – Ethiopia
  - Develop fine-grained Amharic tagsets
  - Incorporating the unsupervised POS tags as features

Amharic text corpora [1/2]

- From tweeter using tweeter API
  - Download tweets everyday
  - 713683 tweets in three years

Year	Users per week	Tweet per week
2014	896	9438
2015	1234	6330
2016	1675	9438

Amharic tweet trends

Amharic text corpora [2/2]

- Collecting Amharic texts from the web using focused crawler
  - Around 1 million clean sentences are collected in one week time
  - Used for:
    - Unsupervised POS tagging feature
    - Spell checking word form preparation

Spell checking

- At **CIDLeS Summer School 2014**: Coding for Language Communities
  - Corpus based
  - Dictionary preparation– corpus frequency list
  - About 500,000 word forms
  - Moderate affix file preparation
  - Around 88% coverage on random web text

The main bottlenecks of Amharic NLP

- Lack of dedicated NLP research group
  - No cooperation or communication between researchers across different institutes
- Lack of Amharic keyboard (or it is complex)
  - Most social media users use the transliteration in stead of the Amharic script (Fidel)
- Most research are done as a master or doctoral thesis and
  - No continuity on the same topic
  - Resources are not available open source

Tasks and collaborations [1/2]

- Amharic tokenizer / segmenter (compound splitting, Sentence demarcation…)
- Pre-processing and normalization approaches
- Development of different corpora(annotation, crawling, cleaning…)
- Standardizing existing NLP tools
  - Guidelines and documentations
  - Open source and free
- Collaborate with different researchers and educators working on Amharic NLP
- Something else ???

Tasks and collaborations [2/2]

- LT group can
  - Help adapting existing NLP tools and approaches to Amharic
  - Give technical help for Master and PHD students working on Amharic NLP

Thank you