EMNLP 2016

Word-Level Language Identification and Predicting Codeswitch Points in Swahili-English Language Data

Mario Piergallini, Rouzbeh Shirvani

Gauri Gautam, Mohamed Chouikha

Why switch to English? Does he not know the Swahili word? Or is it for emphasis? Or does it show his identity as an educated Kenyan?

Why switch to English? Is it for effect? Does it show cultural hipness? Is he copying his friend's behavior, and so showing their shared identity or social closeness?

Text	manze	niko	na	unenge	ile	deadly	leo	tunamanga	nini
Identify languages	Swahili	Sw	Sw	Sw	Sw	English	Sw	Sw	Sw
Segment switch points	manze niko na unenge ile					deadly	leo tunamanga nini		
	Sw					Sw→ En	En→ Sw		
Grammatical analysis	Identify parts of speech and semantic categories of verbs and nouns. Use these to determine how unexpected are these switches? How common are the words?								
Interpretation	What purpose could these switches serve, given the context in the conversation and the participants? Does it indicate something about their identities or relationships?								

Data: Kenyan Interviews

- Conducted by students and professors at a university in Kenya
- Transcribed and labeled for language by Swahili-speaking students at Howard
- Named entities are labeled based on context

- Statistics:
  - 32 interviews
  - 10,105 utterances
  - 188,188 words

  - 84.5% English
  - 15.4% Swahili
  - <0.1% mixed
  - <0.1% other

Data: JamiiForums

- Large, Tanzania-based web forum
- URLs, email addresses, embedded images and emojis were removed
- 22,592 words were labeled for language by a human annotator

- Statistics:
  - 220,434 posts
  - 16,176,057 words
  - 45.8% English
  - 54.1% Swahili
  - <0.1% mixed
  - <0.1% other

Word-Level Language Identification

- Most previous work on language identification operated at the document level
- The first iteration of this workshop increased the attention given to this problem with the shared task and other papers concerning it
- Best performance in previous work ranged from around 78% to 97% accurate, varying based on language pair and whether the test data was from the same domain

Word-Level Language Identification

- Character *n*-grams (1-3), including prefixes and suffixes
- Capitalization feature
- These were used to train a model using logistic regression
- Label probabilities of the preceding and following words were added to the feature vector, which was used to train the final model

Word-Level Language Identification

Train / Test Set:		Interview 10-fold CV		Interview JF Small (6,118 words)		Interview + JF Small JF Large (16,475 words)	
Context Label Prob.?		None	Word ±1	None	Word ±1	None	Word ±1
English	Precision	94.2%	99.4%	41.6%	87.6%	90.1%	99.2%
	Recall	99.0%	99.7%	95.9%	96.6%	96.5%	98.8%
	F1 Score	96.5%	99.5%	58.0%	91.9%	93.2%	99.0%
Swahili	Precision	92.1%	97.9%	98.1%	99.0%	83.7%	95.3%
	Recall	67.0%	97.2%	62.4%	96.2%	64.1%	97.7%
	F1 Score	77.6%	97.5%	76.3%	97.6%	72.6%	96.5%
Accuracy		94.0%	99.3%	69.7%	96.5%	89.0%	98.4%
Cohen's Kappa		0.74	0.98	0.40	0.92	0.66	0.96

Codeswitch Point Prediction

- Solorio and Liu (2008)
  - English-Spanish, spoken conversation
  - Phrase constituent position, POS tags, language of words
- Papalexakis, Nguyen and Doğruöz (2014)
  - Turkish-Dutch, internet forum
  - Language of the word and previous two words, previous CS, emoticons, multi-word expressions
- Approximately 4.5% of words occurred at codeswitch points in the interview data, 5.7% in the JamiiForums data

Codeswitch Point Prediction

- Features:
  - Language of the word and two previous words
  - Whether a codeswitch occurred earlier in the utterance or post
  - We also tried marking the previous words by whether the language was the same or different
  - Count or logarithm of previous words in same or different language
  - Percentage of same language words previously in the utterance or post

Codeswitch Point Prediction

Data:	Interviews		JamiiForums	
Split:	Unbalanced	Balanced	Unbalanced	Balanced
Precision	28.5%	78.3%	27.4%	81.4%
Recall	52.2%	72.6%	51.3%	58.1%
F1 Score	36.8%	75.3%	35.7%	67.8%
Accuracy	97.5%	74.4%	96.9%	67.4%
Cohen's Kappa	0.33	0.52	0.31	0.45

-Model trained using Naïve Bayes
-Highest performance (using unbalanced data) in Solorio & Liu (2008) was F1 of 28%
-Performance (using balanced data) in Papalexakis et al (2014) was F1 of 71% to 77%

Conclusions and Future Directions

- Performance on language identification was quite high, outperforming previous attempts.
  - Accuracy was very high within domain
  - Simplicity may improve results across domains
  - But it may also be that the Swahili-English language pair is more easily distinguished

- The model used in this paper was the basis for the model in our shared task submission

Conclusions and Future Directions

- Performance on codeswitch point prediction was in line with previous attempts
  - Given that codeswitching is never a forced choice, the ceiling for predicting codeswitching using only lower level features may be quite low
  - Improvements may be possible by leveraging conversational structure: codeswitching in the current turn may be more accurately predicted given knowledge of codeswitching behavior in the previous turns

Questions and Discussion