PROCESSING LOW-RESOURCE LANGUAGES: A CASE OF THE YORUBA LANGUAGE

Idris Abdulhameed Oyebode

OUTLINE

- What is low-resource NLP?
- Why is it hard?
- Why are we interested in this?
- Yoruba language
- Part of Speech Tagging for low-resource NLP
- Evaluation
- Conclusion

WHAT IS LOW-RESOURCE NLP?

- Languages that lack large enough monolingual and/or parallel corpora with adequate linguistic resources for developing Natural Language Processing applications.
- Resources include parts-of-speech tags, syntactic features, semantic features.
- There are thousands of them. They include Yoruba, Warlpiri (Pama-Nyungan, Australia), Komi Zyrian (Uralic-Permic, Russia).
- Resource rich are the opposite.
- Examples are English, French, Spanish.
- In between is Medium resource. Includes Czech, Hindi, Hebrew.

WHY IS LOW-RESOURCE NLP HARD?

- The same reason NLP generally is hard.
- Ambiguities at different levels.
  - Parts of speech: bank (Noun or Verb?)
  - Syntax: I saw the lecturer with a telescope. Who is with the telescope?
  - Semantics: I saw him duck.
- Linguistic diversity
  - Words
  - Morphology
  - Parts of speech
  - Language family

WHY ARE WE INTERESTED IN THIS?

- The world is a linguistically diverse place. 6-7k languages.
- Half of them are only spoken and not written. (Lewis, 2009)
- Language consists of many structures
  - sounds, words, morphology, part of speech, syntax, semantics and discourse.
- In the production and understanding of language, humans easily and almost effortlessly assimilate all these structures.
- The objective of NLP is to achieve at least this also, albeit with a computer.
- Core technologies needed to perform this task.
  - Language modelling, Part-of-Speech tagging et cetera.

OBJECTIVES

- Primary
  - Bootstrapping a part-of-speech tagger from the limited available annotated data.
  - Training a Hidden Markov Model with a training data.
  - Decoding (Testing) the Hidden Markov Model on a test data.
- Secondary
  - Implementing Cross-lingual transfer learning with resource rich languages as source languages.
  - Training a Hidden Markov Model using unsupervised learning.

YORUBA LANGUAGE

- Benue-Congo subclass of the Niger-Congo family of languages (Adeniyi, 2007).
- Predominantly spoken language in West Africa with over 40 million native speakers.
- Nigeria and Republic of Benin are major speakers.
- Trinidad and Tobago, Brazil, Cuba and some parts of Europe (Fabunmi, 2005).
- More than 12 dialects which includes Ègbá, Òyó, Òwó, Ìjèbú, Ìjèsà et cetera.
- Yoruba Ajumolo (Standard Yoruba)- dialect which is understood, spoken across cultures and in formal settings.

YORUBA LANGUAGE

- Morphology: derivational
- Compounding:
  - gbale + gbale = gbalegbale
  - sweep + sweep = cleaner
- Affixation:
  - i + to = ito
  - PREFIX + to take care = saliva
- Reduplication:
  - da + oju = dajudaju
  - clear + eye = certainly

PART OF SPEECH TAGGING

- The Data
  - Universal Dependencies (UD) project treebank (Nivre et al., 2018).
  - Has standard of 15 to 17 tags.
  - 2,664 tokens/words and 100 sentences.
  - This is a small dataset.
- The Baseline
- Supervised POS Tagging with Hidden Markov Models
- Cross-lingual Transfer Learning with HMMs

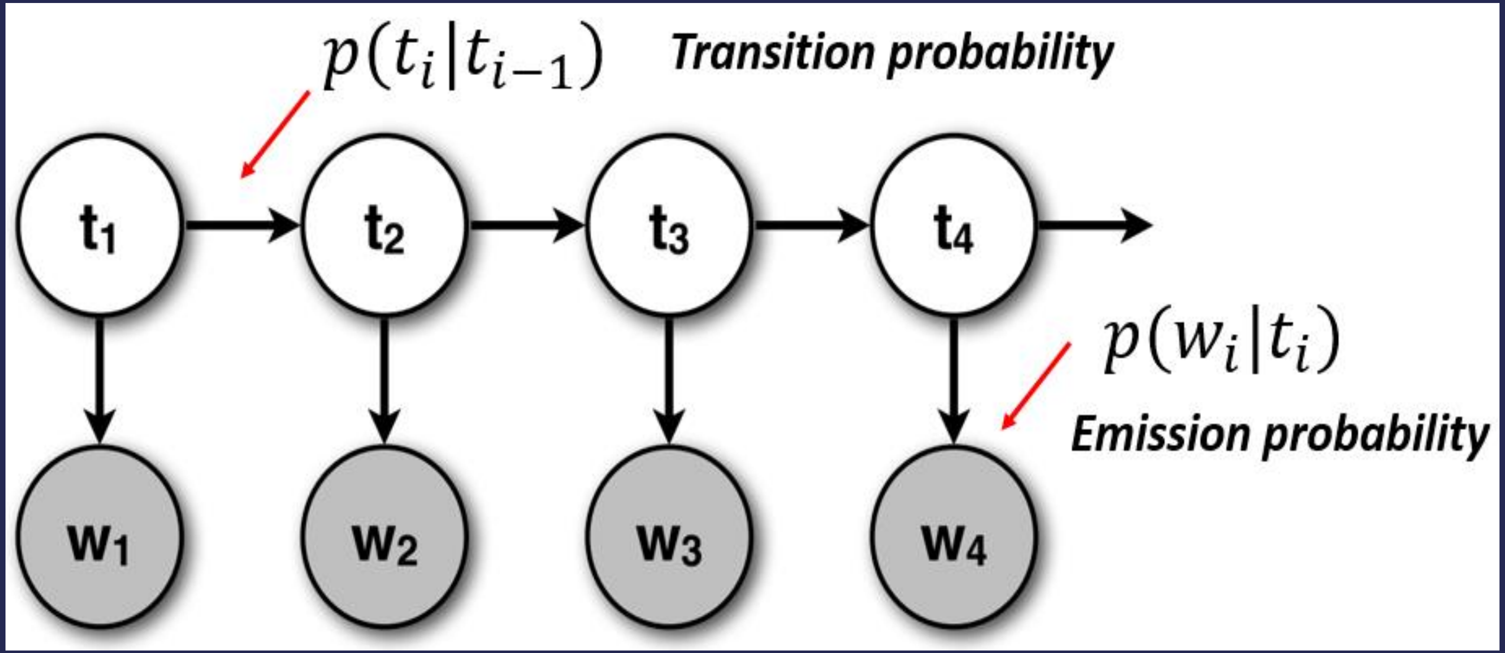
POS tags	Meaning
ADP	Adposition
ADJ	Adjective
AUX	Auxiliary
ADV	Adverb
CCONJ	Coordinating Conjunction
DET	Determiner
INTJ	Interjection
NOUN	Noun
NUM	Numeral
PART	Particle
PROPN	Proper Noun
PRON	Pronoun
PUNCT	Punctuation
SCONJ	Subordinating Conjunction
SYM	Symbol
VERB	Verb
X	Other

APPROACHES

- Baseline:
  - Tagging each word in a sentence/observation with its most frequent/likely tag.
  - $P(tag_n \mid word_1, word_2 .... word_n) \approx P(tag_n \mid word_n)$
  - $P(tag \mid word) = \frac{P(word, tag)}{P(word)}$
- Supervised POS Tagging with Hidden Markov Models
  - generative sequence classifiers that assign labels or classes to units of a sequence.

$$\hat{t}_1^n \approx \underset{\hat{t}_1^n}{argmax} \left(\prod_{k=1}^{n} P(w_k \mid t_k) P(t_k \mid t_{k-1}) \right) P(\langle /s \rangle \mid t_n)$$

HIDDEN MARKOV MODELS



$$P(s,w) = \left(\prod_{k=1}^{n} P\,(w_k \mid s_k)\,P(s_k \mid s_{k-1})\right)$$

Emission Probability

$$P(w_k \mid t_k) = \frac{count\ (t_k, w_k\ )}{count\ (t_k)}$$

Transition Probability

$$P(t_k \mid t_{k-1}) = \frac{count\ (t_{k-1} t_k)}{count\ (t_{k-1})}$$

APPROACHES

- Cross-lingual Transfer Learning with HMMs
  - Transfer learning has been the most frequently used approach to low-resource Natural Language Processing.
  - Useful when there is a very small amount of annotated data or there is no annotated data at all.
  - Many Natural Languages have been found to share similar properties.
    - parts of speech, word order, syntax et cetera.
  - Transfer only our transition probabilities model from resource rich languages (9 languages) to the Yoruba language.

$$\hat{t}_1^n \approx \underset{\hat{t}_1^n}{argmax} \left(\prod_{k=1}^{n} P_{LR}(w_k \mid t_k)\, P_{RR}(t_k \mid t_{k-1}) \right) P_{RR}(\langle/s\rangle \mid t_n)$$

SMOOTHING

- Problem of data sparsity.
- A common problem in speech and language processing.
- Witten-Bell Smoothing (Witten and Bell, 1991) :
  - Use the higher order model (bigram in this case) if the bigram was seen in the training data, otherwise, we back off to the unigram model .

$$P_{WB}\left(w_k \mid w_{k-n+1}^{k-1}\right) = \lambda_{t_{k-n+1}^{k-1}} P_{ML}\left(w_k \mid w_{k-n+1}^{k-1}\right) + \left(1 - \lambda_{t_{k-n+1}^{k-1}}\right) P_{WB}\left(w_k \mid w_{k-n+2}^{k-1}\right)$$

TESTING

- Machine Learning classification problem.
- Decoding
  - Viterbi algorithm
  - memorized and iterative solution

$$\delta_k(s) = \max_{s_0 \ldots s_{k-1} s} P(s_0 \ldots s_{k-1}\, s, w_1 \ldots w_{i-1})$$

which for HMMs becomes:

$$\delta_k(s) = \max_{s'} P(s \mid s') P(w_k \mid s)\ \delta_{k-1}(s')$$

  - Back pointers to keep track of the probabilities at each step:

$$\varphi_k(s) = \underset{s'}{argmax}\, P(s \mid s') P(w_k \mid s)\ \delta_{k-1}(s')$$

EVALUATION

- Data size:

Data Set	Tokens	Unknown
Training	1,748	
Test	928	188 (20.26%)

- Baseline: with VERB for unknowns

Folds	Accuracy (%) (Lowercase)	Accuracy (%) (Sentence case)
Fold 1	76.31	76.31
Fold 2	83.99	88.35
Fold 3	83.85	87.11
Fold 4	78.76	79.36
Mean	**80.73**	**82.78**

EVALUATION

- Supervised HMM Results

Folds	Accuracy (%) (Lowercase)	Accuracy (%) (Sentence case)
Fold 1	78.38	79.17
Fold 2	87.04	86.75
Fold 3	86.13	86.13
Fold 4	83.58	82.40
Mean	**83.78**	**84.64**

EVALUATION

- Cross-lingual Transfer Learning Results

	Accuracy (%)								
Source	English	Catalan	Spanish	French	Swedish	Portuguese	Naija	Danish	Italian
Fold 1	85.15	84.34	83.84	82.38	81.40	83.36	84.67	84.39	83.69
Fold 2	84.86	85.59	85.00	83.11	81.22	85.44	86.17	84.71	85.88
Fold 3	81.02	79.21	79.37	80.57	79.37	78.77	80.72	79.22	77.86
Fold 4	80.60	78.54	78.86	77.74	78.37	80.45	76.15	80.76	75.67
Mean	**82.91**	**81.92**	**81.76**	**80.95**	**80.09**	**82.05**	**81.93**	**82.27**	**80.78**

EVALUATION

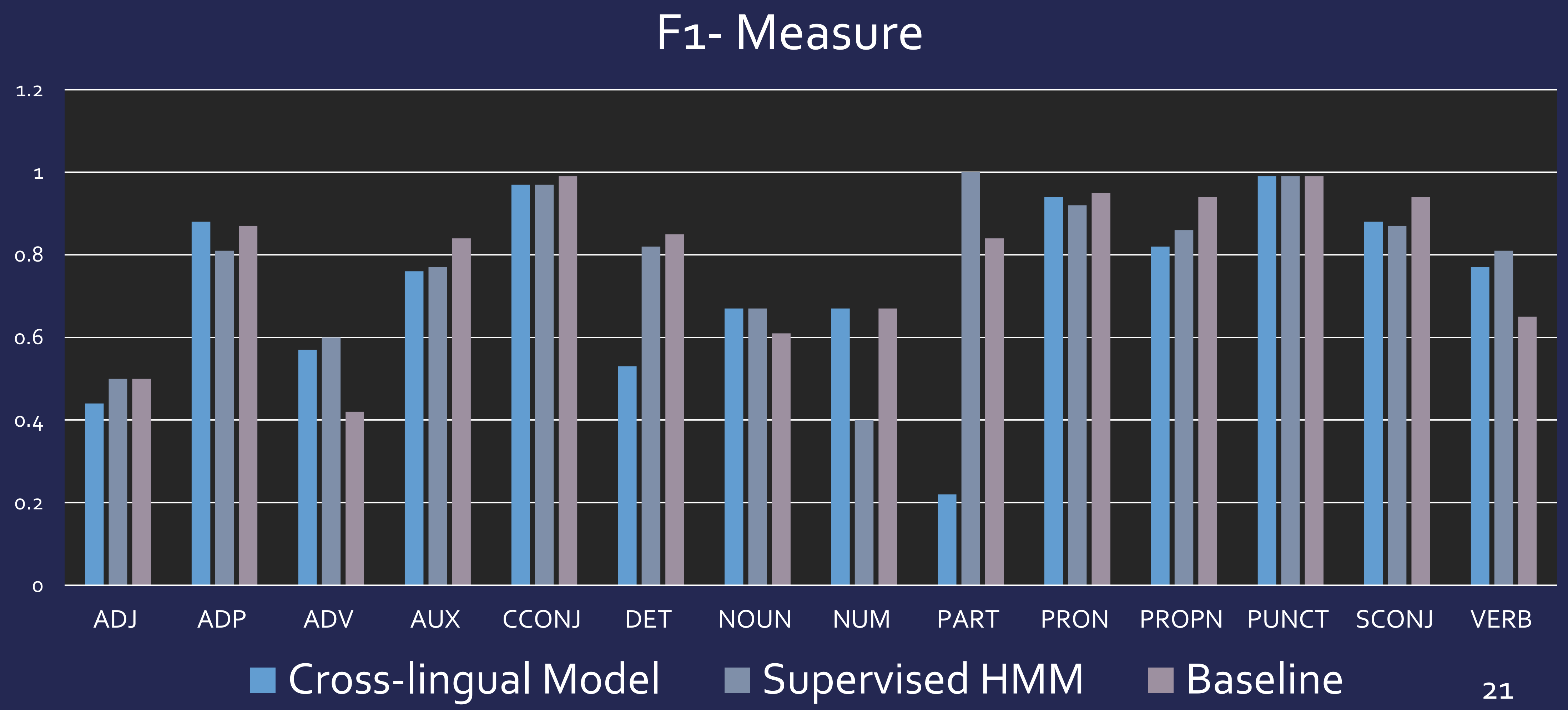
- A model with naïve transition probabilities

Folds	Accuracy (%) (Lowercase)	Accuracy (%) (Sentence case)
Fold 1	74.69	75.90
Fold 2	72.81	72.81
Fold 3	79.62	83.99
Fold 4	83.69	84.02
Mean	**77.70**	**79.18**

SUMMARY OF MODELS

Models	Baseline	Supervised HMM	Naïve Transition HMM	Cross-lingual Model
Accuracy (%)	82.78	84.64	79.18	82.91

SUMMARY OF MODELS



BEST MODEL- CONFUSION MATRIX

Tags	ADP	ADV	AUX	CCONJ	NOUN	PRON	PROPN	PUNCT	SCONJ	VERB
ADP	5.3	0	0.3	0	0	0	0	0	0	0.1
ADV	0.2	2.7	0.2	0	1.2	0.1	0	0	0.2	0.1
AUX	0.1	0.1	5.4	0	0	0	0	0	0	0.5
CCONJ	0	0.1	0	3.8	0	0	0	0	0	0
NOUN	0.4	0	0	0	5.1	0	0	0	0	0.4
PRON	0.2	0	0.2	0	0.3	18.5	0	0	0	0.9
PROPN	0.1	0	0	0	0.8	0	3.8	0	0	0.1
PUNCT	0	0.1	0	0	0.2	0	0	18.2	0	0
SCONJ	0	0.4	1.0	0	0	0	0	0	5.8	0.1
VERB	0.9	0.3	0.8	0	1.2	1.1	0	0	0	14.0

CONCLUSION AND FUTURE WORK

- We have implemented 3 models for supervised POS Tagging of the Yoruba language.
- We discovered that using Hidden Markov Models with lowercased words was a poorer feature as against being a richer feature in the English Language.
- We also discovered transferring transition probabilities from very related resource rich languages has significant effect on low-resource languages.
- In the future, we hope to perform POS projections with word alignment and unsupervised training for the Yoruba language.
- Following this, we hope to implement POS tagging for a now larger annotated corpus of Yoruba using Neural networks and Bi-LSTMs.
- We also hope to extend this to other low-resource languages and further this processing to other aspects of NLP.

REFERENCES

- Adeniyi, H. R. (2007). A comparative study of reduplication in Edo and Yorùbá. *MorphOn:e-journal of morphology*, pages 1–23.
- Fabunmi, F. A. & Salawu, A. S. (2005). Is Yorùbá an endangered language? *Nordic Journal of African Studies*, 14(3):391–408.
- Lewis, P. (2009).Ethnologue: Languages of the World. SIL International.
- Nivre, Joakim; Abrams, Mitchell; Agić, Željko; et al., (2018). Universal Dependencies 2.3, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Witten, I. H. & Bell, T. C. (1991). The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression. *IEEE Transactions on Information Theory*, *37*(4), 1085–1094.


THANK
YOU
MAAKE
TERMA KASIH
RAIBH MAITH AGAT
GRAZIE
MULTUMESC
JUSPAXAR
OBRIGADO
MATONDO
SALAMAT
KIITOS
MOCHCHAKKERAM
MERCI
KIA ORA
MULTUMESC
CHOKRANE
SALAMAT
CAM ON BAN
RAIBH MAITH AGAT
MERCI
MERCI
MOCHCHAKKERAM
OBRIGADO
MOCHCHAKKERAM
GRAZIE
ASANTE
CHOKRANE
MATUR NUWUN
UA TSAUG RAU KOJ
MOCHCHAKKERAM
MATONDO
WELALIN
KIITOS
CHOKRANE
UA TSAUG RAU KOJ
DANK JE
OBRIGADO
DANK JE
RAIBH MAITH AGAT
SPASIBO
VINAKA
NIRRINGRAZZJAK
MAMANA
SPASIBO
ARIGATO
KIITOS
MAAKE
DANKON
MULTUMESC
OBRIGADO
MOCHCHAKKERAM
NIRRINGRAZZJAK
MOCHCHAKKERAM
OBRIGADO
SPASIBO