Language Model Based Arabic Word Segmentation

By Saleh Al-Zaid
Software Engineering

Content

- Introduction
- Morpheme Segmentation
- New Stems Acquisition
- Performance Evaluation
- Conclusion

Introduction

- **Paper:** Language Model Based Arabic Word Segmentation, 2003.
- **By:** Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam and Hany Hassan .
- **Objective:** introduce an algorithm to segment and acquire new Arabic stems form un-segmented Arabic corpus.

Segmentation?

- Divide words into parts (morphemes) using the pattern: prefix*-stem-suffix*
- * = zero or more
- Prefix indicated by #
- Suffix indicated by +
- Example: السيارة = ة+ سيار #ال

Algorithm Implementation

1. Language model training on a manually morpheme-segmented small corpus (form 20K to 100K words).

2. Segmentation of input text into a sequence of morphemes using the language model parameters.

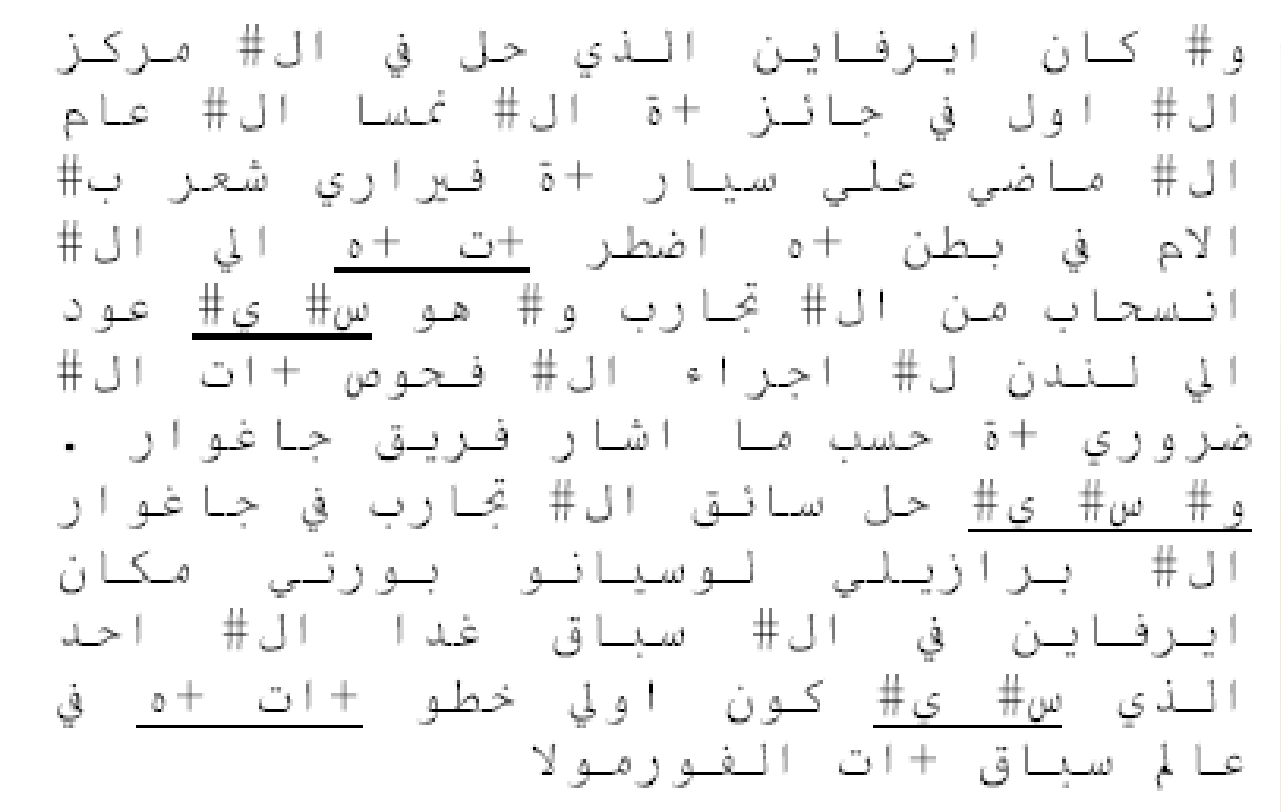
3. Acquisition of new stems from a large unsegmented corpus.

Morpheme Segmentation

1. Trigram Language Model
2. Decoder for Morpheme Segmentation
   - **Possible Segmentations of a Word**

Trigram Language Model

- Sample of manually segmented corpus



Trigram Language Model

- Buckwalter equivalence in English:

w# kAn AyrfAyn Al*y Hl fy Al# mrkz Al# Awl fy jA}z +p Al# nmsA Al# EAm Al# mADy Ely syAr +p fyrAry $Er b# AlAm fy bTn +h ADTr **+t +h** Aly Al# AnsHAb mn Al# tjArb w# hw **s# y#** Ewd Aly lndn l# AjrA' Al# fHwS +At Al# Drwry +p Hsb mA A$Ar fryq

- Trigram: $p(m_i \mid m_{i-1}, m_{i-2})$

Table of Segments

Words		Prefixes		Stems		Suffixes	
Arabic	Translit.	Arabic	Translit.	Arabic	Translit.	Arabic	Translit.
الولايات	*AlwlAyAt*	ال#	*Al#*	ولاي	*wlAy*	+ ات	*+At*
حياته	*HyAth*			حيا	*HyA*	+ ت + ه	*+t +h*
للحصول	*llHSwl*	ل# ال#	*l# Al#*	حصول	*HSwl*		
الى	*AlY*			الى	*AlY*		

Table 1 Segmentation of Arabic Words into Prefix*-Stem-Suffix*

To be used in the algorithm next steps

Decoder for Morpheme Segmentation

- Goal: to find the morpheme sequence which maximizes the trigram probability of the input sentence, as in:

SEGMENTATIONbest = *Argmax* $\prod_{i=1,N} p(m_i | m_{i-1}\, m_{i-2})$

N = number of morphemes in the input

Possible Segmentations of a Word

- Depends on the table of Prefix/Suffix

Prefixes		Suffixes	
ال	ال#	ات	+ات
بـال	ب# ال#	اتهـا	+ات +ها
وبـال	و# ب# ال#	ونـهم	+ون +هم

Table 2 Prefix/Suffix Table

- Each new token is assumed to have: *prefix*-stem-suffix** structure and compared against prefix/suffix table.

Possible Segmentations of a Word

- Steps to find possibilities:

i) Identify all of the matching prefixes and suffixes from the table,

ii) Further segment each matching prefix/suffix at each character position, and

iii) Enumerate all *prefix*-stem-suffix** sequences derivable from (i) and (ii).

Possible Segmentations of a Word

- Example: suppose "و اكررها" is new token which is : *wAkrrhA*,

- Using :

  SEGMENTATIONbest = *Argmax* IIi=1,N p(*mi*|*mi-1 mi-2*),

  the possible segmentations are:

Possible Segmentations of a Word

	Prefix	Stem	Suffix	Seg Scores
S1	∅	*wAkrrhA*	∅	2.6071e-05
S2	∅	*wAkrrh*	+*A*	1.36561e-06
S3	∅	*wAkrr*	+*hA*	9.45933e-07
S4	*w#*	*AkrrhA*	∅	2.72648e-06
S5	*w#*	*Akrrh*	+*A*	5.64843e-07
S6	*w#*	*Akrr*	+*hA*	4.52229e-05
S7	*wA#*	*krrhA*	∅	7.58256e-10
S8	*wA#*	*krrh*	+*A*	5.09988e-11
S9	*wA#*	*krr*	+*hA*	1.91774e-08
S10	*w# A#*	*krrhA*	∅	7.69038e-07
S11	*w# A#*	*krrh*	+*A*	1.82663e-07
S12	***w# A#***	***krr***	***+hA***	**0.000944511**

Table 3 Possible Segmentations of و ا كـ ر ر هـا (*wAkrrhA*)

Acquisition of New Stems

- Form large un-segmented corpuses.
- Follow this process:

- **Initialization:** Develop the seed segmenter Segmenter0 trained on the manually segmented corpus Corpus0, using the language model vocabulary, Vocab0, acquired from Corpus0.

Acquisition of New Stems

- **Iteration:** For i = 1 to N, N = the number of partitions of the unsegmented corpus:
- i. Use Segmenter$_{i-1}$ to segment Corpus$_i$.
- ii. Acquire new stems from the newly segmented Corpus$_i$. Add the new stems to Vocab$_{i-1}$, creating an expanded vocabulary Vocab$_i$.
- iii. Develop Segmenter$_i$ trained on Corpus0 Corpus$_i$ with Vocab$_i$.

Performance Evaluation

- Using the formula:
- E = (*number of incorrectly segmented tokens / total number of tokens*) *x 100*
- Example: ليتر : ل# ي# تر
- For the paper, the evaluation is done in test corpus with 28,449 words by using 4 manually segmented seed corpora with 10K, 20K, 40K, and 110K words.

Results

Manually Segmented Training Corpus Size	Unknown Stem	Other Errors	Total # of Errors
10 K Words	1,844 (76.9%)	455 (19.0%)	2,397
20 K Words	1,174 (71.1%)	395 (23.9%)	1,651
40 K Words	1,005 (69.9%)	351 (24.4%)	1,437
110 K Words	333 (39.6%)	426 (50.7%)	841

Table 7 Segmentation Error Analyses

Conclusion

- Language Model Based Arabic Word Segmentation Algorithm can segment and acquire new stems as long we have good training manually segmented corpus
- It give more good results as the training corpus have large number of stems

Thank you

Q & A