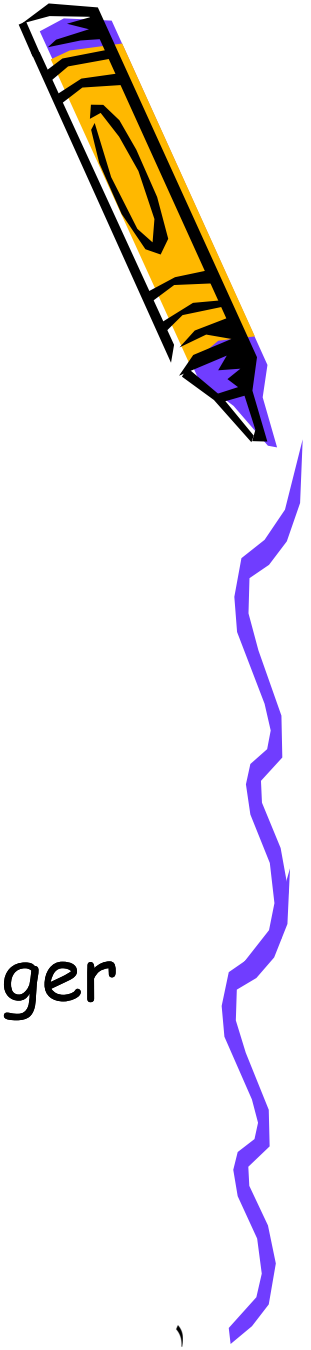


A Hidden Markov Model-Based POS Tagger for Arabic

ICS 482 Presentation

A Hidden Markov Model- Based POS Tagger for Arabic

By

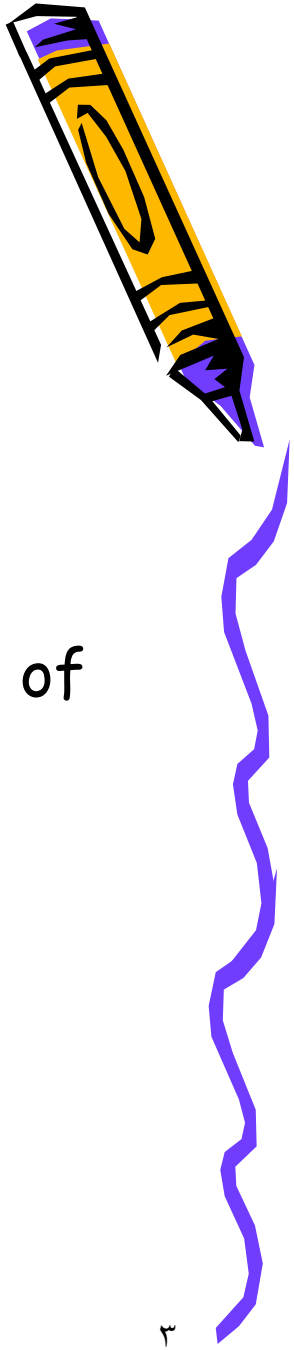
Saleh Yousef Al-Hudail

222154

OUTLINE

- Introduction
- Arabic Lexical Characteristics and POS Tag Set Description
  - Nouns, Pronouns, Verbs, Particles
- The HMM-based POS Tagger
  - Approach
    - The Tokenizer
    - The Stemmer
    - The POS Tagger
  - Construction of the HMM Model
- Summary

About the Paper



- Written by Fatma Al Shamsi and Ahmed Guessoum. (2006).
- Department of Computer Science – University of Sharjah in UAE.

Introduction

- Purpose:
  - Arabic language is spoken by over 300 million people.
  - NLP for Arabic is yet to achieve the aimed quality and robustness levels.
- Many words in Arabic can have the same constituent letters but different pronunciations, thus, presence of diacritics:
  - fatHa, Dhamma, kasra, sukuun.
- Absence of these is very common in Standard Arabic. Adds a lot of lexical ambiguity.
- Contextual vs. lexical !!

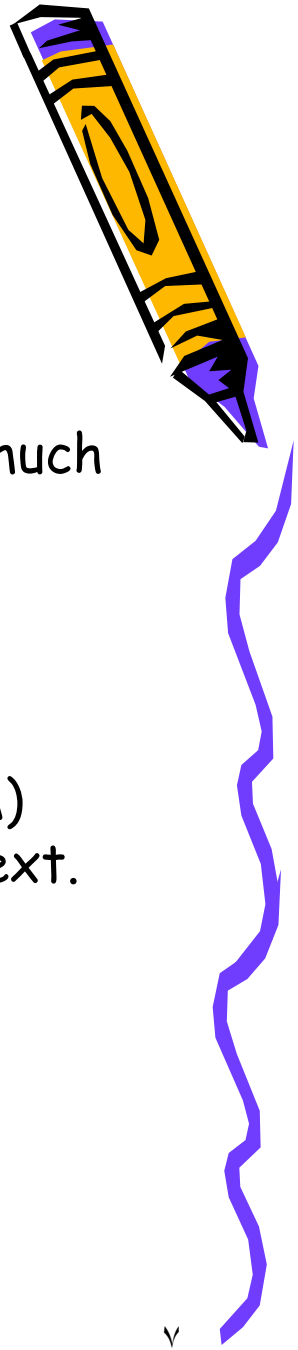
POS Tagging Definition

- POS tagging is the process of assigning a part-of-speech tag such as noun, verb, pronoun, preposition, adverb, adjective or other tags to each word in a sentence (Jurafsky and Martin, 2000).
- Based on the context to resolve lexical ambiguity.
- Two approaches of POS taggers: rule based and trained ones.

Why HMM Model??

- HMM Model make use of previous events to assess the probability of the current events, i.e., N-gram.
- HMM is superior to other models with regards to training speed.
- Hence is suitable for application with large amount of data to be processed.

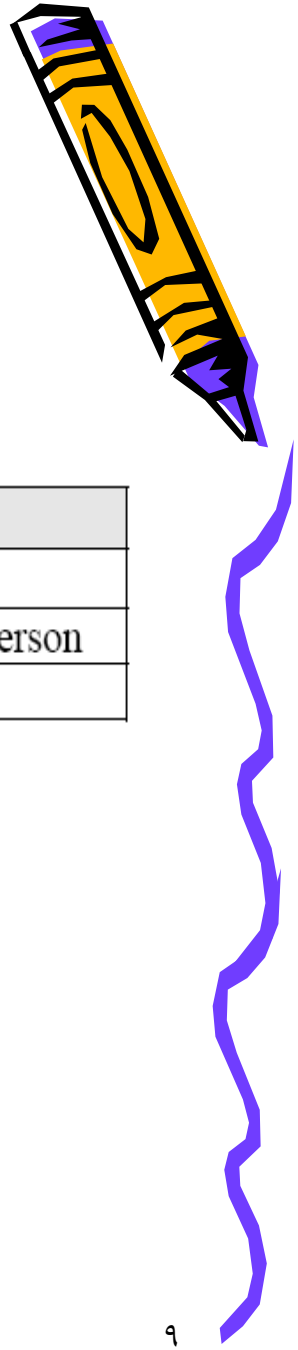
Duh & Kirchhoff(DK) vs. this paper



- Since Arabic is rich in morphology and most POS as available as inflections or affixes, there has not been much work done in Arabic Tagging.
- **Performance:** 68.48% vs. 97%
- **Methodology:** similar to Support Vector Machine (SVM) uses Linguistic Data Consortium (LCD) vs. raw Arabic text.

Lexical Characteristics and POS Tag Set Description

- **Selection criteria of tag set:**
  - Ensure that the tag set is rich enough to allow a good training and a good performance of the HMM-based POS tagger.
  - The tag set is small enough to make the training of the POS tagger computationally feasible.
- **Description of POS Tag Set:**
  - Two Gender masculine and feminine (F, M).
  - Three persons speaker (first person), the person being addressed (second person), the person that is not present (third person). As (1, 2, 3).
  - Three numbers (S, D, P).



	Suffix	Stem	Prefix
Arabic	ين (iyn)	أكل ('kul)	ت (ta)
Morphological Analysis	Suffix, 2nd person, feminine	Verb	Prefix, 2nd person
Meaning	You feminine	eat	you

Table 2 : Morphological Structure of ين ت أكل *(ta'kuliyna/you eat)*

Description of POS Tag Set
Continued…

- ***Nouns***
  - Arabic nouns can be subcategorized into adjectives, proper nouns and pronouns. A noun can be definite or indefinite.
    NOUN (noun), ADJ (adjective), PNOUN (proper noun), PRON (pronoun), INDEF (indefinite noun), DEF(definite noun).

  - There are three grammatical cases in Arabic : the nominative (الرفع), the accusative (النصب) and the genitive (الجر). These cases are distinguished based on the noun **suffixes (SUFF)**.

Case	Nominative	Genitive	Accusative	All
Word	مسـلمون	مسـلمين	مسـلمان	مسـلمات
Transliteration	muslimuwn	muslimiyn	muslimaan	muslimaat
Meaning	Muslims (masc., plural)	Muslims (masc., plural)	Two Muslims (masc., dual)	Muslims (fem., plural)
Suffix POS tag	ون/ SUFF_M_P	يـن/ SUFF_SUBJ_ALL	ان/ SUFF_M_D	ات/SUFF_F_P

Table 3 : Different plural and dual forms of the word مسـلم (muslim)

Description of POS Tag Set Continued…

- ***Pronouns***
  - We have selected to tag demonstrative, possessive and direct object pronouns with the following tags : **DPRON**, **PPRON** and **SUFFDO**

Word	Morphology Analysis	POS Tag
أنـت	Second person singular feminine/masculine pronoun	PRON_2S

Table 6 : Tagging of the pronoun أنت *(you)*

- ***Verbs***
  - PVERB (perfect verb), IVERB (imperfect verb), CVERB (imperative verb), MOOD_SJ (subjunctive or jussive), MOOD_I (indicative), SUFF_SUBJ (suffix subject), FUTURE (future).

Description of POS Tag Set Continued...

- ***Particles***
  - The grammatical function of these words is to come before a noun and change its case from nominative to accusative represented as FUNC_WORD.
  - Include interrogation, conjunction, preposition, and negation particles. As, INTERROGATE, CONJ , PREP and NEGATION.
  - Numeral quantities can be written in two different ways : numerically and alphabetically.

Word	Meaning	POS Tag
الحادي	The first of	DEF+ADJ
عشر	Ten	NOUN
من	From	PREP
أكتوبر	October	PNOUN

  - Numerically can be given a single tag NUM.

POS TAG Set Used

ADJ	EXCEPT	PPRON_2FP	PRON_3D	SUFF_M_P
CONJ	FUNC_WORD	PPRON_3FP	PRON_3FP	SUFF_SUBJ_1P
CVERB	FUTURE	PREP	PRON_3FS	SUFF_SUBJ_2D
DEF	INTERROGATE	PRON	PRON_3MP	SUFF_SUBJ_2FP
DPRON_F	IV1P	PRON_1P	PRON_3MS	SUFF_SUBJ_2MP
DPRON_FD	IV2	PRON_1S	PVERB	SUFF_SUBJ_2S
DPRON_FP	IV3	PRON_2	SHORT_FORM	SUFF_SUBJ_3FD
DPRON_FS	IVERB	PRON_2D	SUFF_F_D	SUFF_SUBJ_ALL
DPRON_MD	NEGATION	PRON_2FP	SUFF_F_P	SUFF_SUBJ_FP
DPRON_MP	NOUN	PRON_2MP	SUFF_F_S	SUFF_SUBJ_MP
DPRON_MS	PNOUN	PRON_2S	SUFF_M_D	SUFF_S_INDEF

The HMM-Based POS Tagger

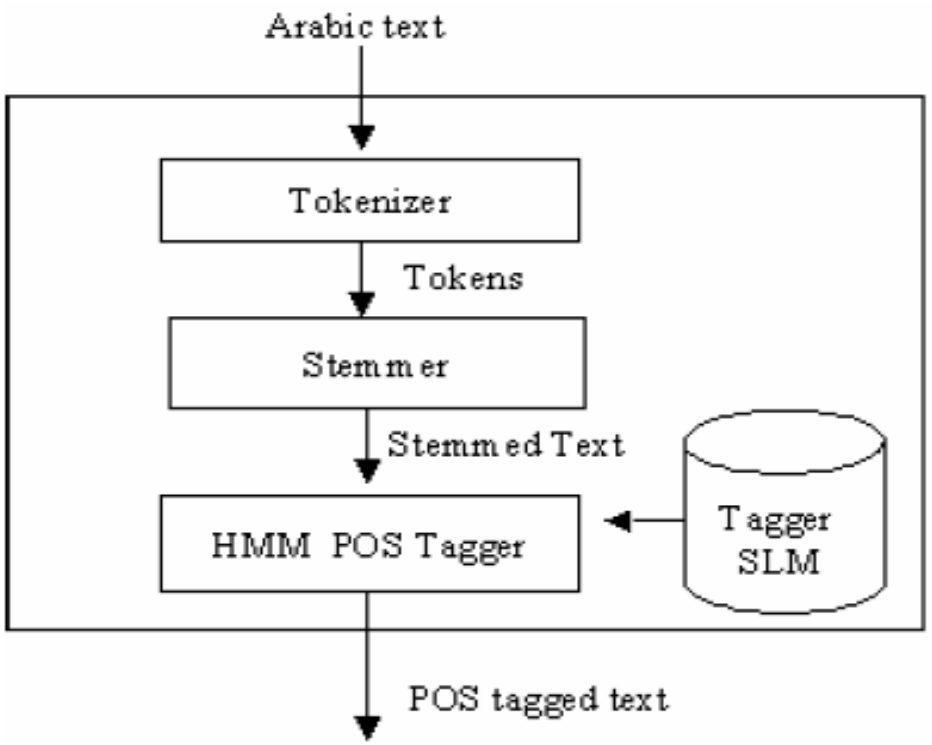


Figure 1 : HMM POS Tagger architecture

Stemmer & Tagger

- The stemmer in (Buckwalter, 2002) returns all valid segmentations as follows:
  - An Arabic prefix length can go from zero to four characters.
  - The stem can consist of one or more characters.
  - And the suffix can consist of zero to six characters.

- The tagger have constructed trigram language models and used the trigram probabilities in building the HMM model, which is expressed by:
  - The set of states S
  - The observation sequence O
  - A matrix A which stores transition probabilities between states (= tag)
  - And matrix B which stores state observation probabilities (called emission probabilities)

The transition probability from the state Noun to the state Adjective is P(ADJ | DEF NOUN) which is formally $P(n_i|n_{i-2}\ n_{i-1})$ and P(كبير |ال DEF بيت NOUN ADJ) is the observation probability that كبير (kabiir / big) is an adjective which is formally $P(w_i|w_{i-2}n_{i-2}\ w_{i-1}n_{i-1}\ n_i)$.

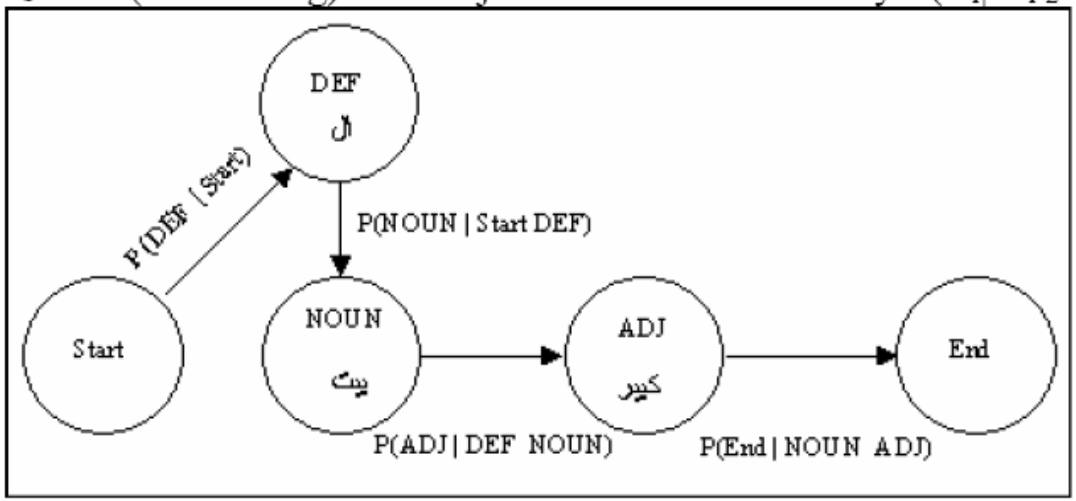


Figure 2 : HMM diagram of the Arabic sentence : البيت كبير (al-baytu kabiir / the house is big)

Constructing the HMM Model

- phrases in Arabic : noun phrase and verb phrase.

1 القمر منير ('al qamaru muniyruN / t
2 السماء صافية ('al samaa'u Saafiyah

- Noun phrase structure expression : [*CONJ *PREP *DEF *FUNC_WORD *[NEGATION INTERROGATE]] [NOUN PNOUN ADJ] [*SUFF% *%PRON%]

- Verb phrase structure expression :
- [*CONJ *PREP *[NEGATION INTERROGATE] *FUTURE *IV%] [PVERB IVERB CVERB] [*SUFF% *%PRON%]

1 أكل الولد التفاحة ('akala 'al-waladu 'al-tuffaaHah / T
2 يأكل حسن التفاحة (ya'kulu Hassan 'al-tuffaaHah

Constructing the HMM Model (contd.)

The trigram DPRON_MS DEF NOUN is 0.459 but the trigram DPRON_MS DEF PVERB is not estimated because it was not seen in the training corpus.

Word	فرنسي	شخص	ال	هذا
Transliteration	faransiyy	shakhS	al	haadhaa
Meaning	French	person	is	This
POS Tag	ADJ	NOUN	DEF	DPRON_MS

Table 9 : POS tagging of sentence : هذا الشخص فرنسي (haadhaa 'alshakhS faransiyy)

Constructing the HMM Model (contd.)

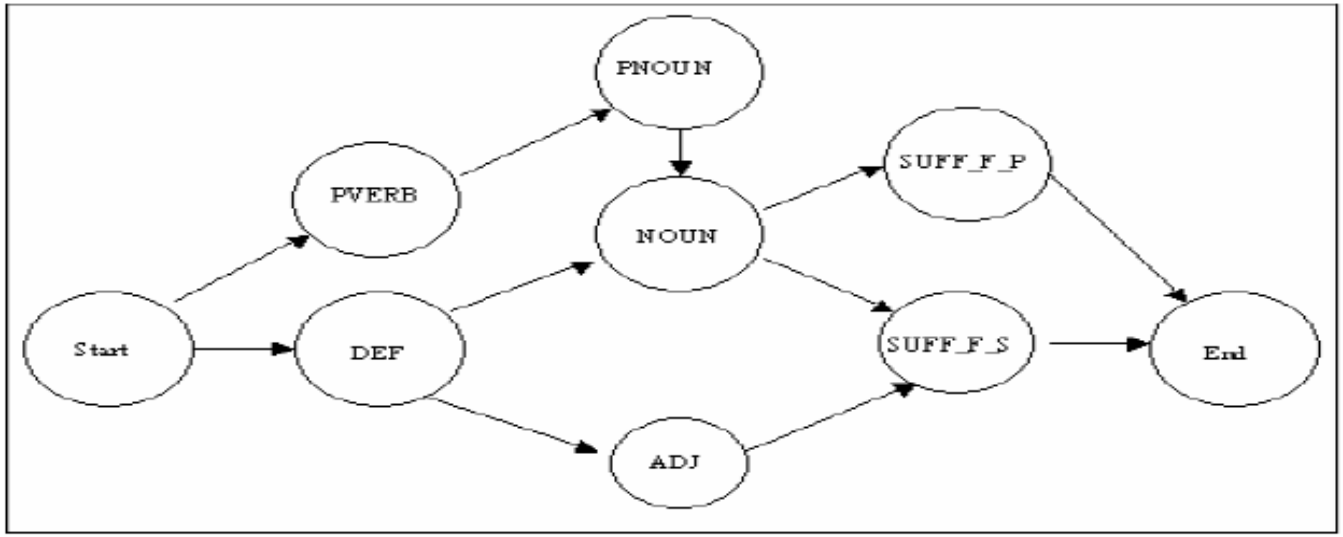
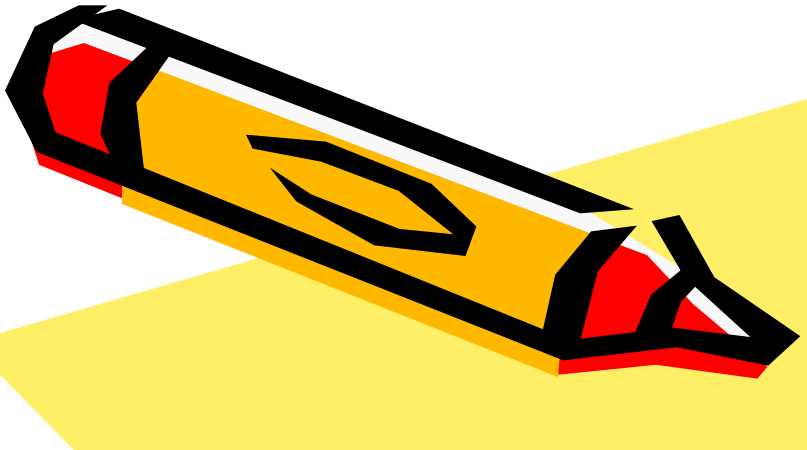


Figure 3 : Partial POS HMM model

Summary

- Have presented a statistical approach that uses HMM to do POS tagging of Arabic text.
- Have analyzed the Arabic language quite systematically and have come up with a good tag set of 55 tags.
- Have then used Buckwalter's stemmer to stem Arabic corpus and we manually corrected any tagging errors.
- Designed and built an HMM-based model of Arabic POS tags.
- One of the greatest advantages of having a trainable POS tagger is that it will speed up the process of tagging huge corpora.



Thank you

If you have any Question
DO NOT hesitate!!