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


• 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).
<table>
<thead>
<tr>
<th>Arabic</th>
<th>Suffix</th>
<th>Stem</th>
<th>Prefix</th>
</tr>
</thead>
<tbody>
<tr>
<td>ت (ta)</td>
<td>ين (iy)</td>
<td>أكل (kul)</td>
<td></td>
</tr>
<tr>
<td>Morphological Analysis</td>
<td>Suffix, 2nd person, feminine</td>
<td>Verb</td>
<td>Prefix, 2nd person</td>
</tr>
<tr>
<td>Meaning</td>
<td>You feminine</td>
<td>eat</td>
<td>you</td>
</tr>
</tbody>
</table>

*Table 2: Morphological Structure ofَ ت أكل َ (ta'kuliyna/you eat)*
• **Nouns**
  - Arabic nouns can be subcategorized into adjectives, proper nouns and pronouns. A noun can be definite or indefinite. **NOUN** (noun), **ADJ** (adjective), **PNNOUN** (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**).
<table>
<thead>
<tr>
<th>Case</th>
<th>Nominative</th>
<th>Genitive</th>
<th>Accusative</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>مسلمون</td>
<td>مسلمین</td>
<td>مسلمان</td>
<td>مسلمات</td>
</tr>
<tr>
<td>Transliteration</td>
<td>muslimuwn</td>
<td>muslimiyn</td>
<td>muslimaan</td>
<td>muslimaat</td>
</tr>
<tr>
<td>Meaning</td>
<td>Muslims (masc., plural)</td>
<td>Muslims (masc., plural)</td>
<td>Two Muslims (masc., dual)</td>
<td>Muslims (fem., plural)</td>
</tr>
<tr>
<td>Suffix POS tag</td>
<td>أن/ SUFF_M_P</td>
<td>بين/ SUFF_SUBJ_ALL</td>
<td>أن/ SUFF_M_D</td>
<td>أت/SUFF_F_P</td>
</tr>
</tbody>
</table>

*Table 3: Different plural and dual forms of the word مسلم (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.

<table>
<thead>
<tr>
<th>Word</th>
<th>Morphology Analysis</th>
<th>POS Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>أنت</td>
<td>Second person singular feminine/masculine pronoun</td>
<td>PRON_2S</td>
</tr>
</tbody>
</table>

*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.

<table>
<thead>
<tr>
<th>Word</th>
<th>Meaning</th>
<th>POS Tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>الحادي</td>
<td>The first of</td>
<td>DEF+ADJ</td>
</tr>
<tr>
<td>عشر</td>
<td>Ten</td>
<td>NOUN</td>
</tr>
<tr>
<td>من</td>
<td>From</td>
<td>PREP</td>
</tr>
<tr>
<td>أكتوبر</td>
<td>October</td>
<td>PNNOUN</td>
</tr>
</tbody>
</table>

- Numerically can be given a single tag NUM.
POS TAG Set Used

<table>
<thead>
<tr>
<th>ADJ</th>
<th>EXCEPT</th>
<th>PPRON_2FP</th>
<th>PRON_3D</th>
<th>SUFF_M_P</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONJ</td>
<td>FUNC_WORD</td>
<td>PPRON_3FP</td>
<td>PRON_3FP</td>
<td>SUFF_SUBJ_1P</td>
</tr>
<tr>
<td>CVL</td>
<td>PREP</td>
<td>PRON_3FS</td>
<td>SUFF_SUBJ_2D</td>
<td></td>
</tr>
<tr>
<td>DEF</td>
<td>INTERROGATE</td>
<td>PRON</td>
<td>PRON_3MP</td>
<td>SUFF_SUBJ_2FP</td>
</tr>
<tr>
<td>DPRON_F</td>
<td>IV1P</td>
<td>PRON_1P</td>
<td>PRON_3MS</td>
<td>SUFF_SUBJ_2MP</td>
</tr>
<tr>
<td>DPRON_FD</td>
<td>IV2</td>
<td>PRON_1S</td>
<td>PVERB</td>
<td>SUFF_SUBJ_2S</td>
</tr>
<tr>
<td>DPRON_FP</td>
<td>IV3</td>
<td>PRON_2</td>
<td>SHORT_FORM</td>
<td>SUFF_SUBJ_3FD</td>
</tr>
<tr>
<td>DPRON_FS</td>
<td>IVERB</td>
<td>PRON_2D</td>
<td>SUFF_F_D</td>
<td>SUFF_SUBJ_ALL</td>
</tr>
<tr>
<td>DPRON_MD</td>
<td>NEGATION</td>
<td>PRON_2FP</td>
<td>SUFF_F_P</td>
<td>SUFF_SUBJ_FP</td>
</tr>
<tr>
<td>DPRON_MP</td>
<td>NOUN</td>
<td>PRON_2MP</td>
<td>SUFF_F_S</td>
<td>SUFF_SUBJ_MP</td>
</tr>
<tr>
<td>DPRON_MS</td>
<td>PNOUN</td>
<td>PRON_2S</td>
<td>SUFF_M_D</td>
<td>SUFF_S_INDEF</td>
</tr>
</tbody>
</table>
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 $P(\text{ADJ} \mid \text{DEF NOUN})$ which is formally $P(n_i \mid n_{i-2} \ n_{i-1})$ and $P(\text{كبیر} \mid \text{ال NOUN ADJ})$ is the observation probability that $\text{كبیر} (\text{kabiir} / \text{big})$ is an adjective which is formally $P(w_i \mid w_{i-2} \ n_{i-2} \ w_{i-1} \ n_{i-1} \ n_i)$.

Figure 2: HMM diagram of the Arabic sentence: $\text{ال البيت كبير} (\text{al-baytu kabiir} / \text{the house is big})$
Constructing the HMM Model

- phrases in Arabic: noun phrase and verb phrase.

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\%]

القرم منير 1 (al qamaru muniyruN / t

السماء صافية 2 (al samaa'u Saafiyah

أكل الولد النافحة 1 ('akala 'al-waladu 'al-tuffaaHah / I

يأكل حسن النافحة 2 (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.

<table>
<thead>
<tr>
<th>Word</th>
<th>فرنسي</th>
<th>شخص</th>
<th>ال</th>
<th>هذا</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transliteration</td>
<td>faransiyy</td>
<td>shakhS</td>
<td>al</td>
<td>haadhaa</td>
</tr>
<tr>
<td>Meaning</td>
<td>French</td>
<td>person</td>
<td>is</td>
<td>This</td>
</tr>
<tr>
<td>POS Tag</td>
<td>ADJ</td>
<td>NOUN</td>
<td>DEF</td>
<td>DPRON_MS</td>
</tr>
</tbody>
</table>

*Table 9: POS tagging of sentence: فرنسي شخص آل هذا (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!!