Meningkatkan metode *Rule-Based* untuk *Part-Of-Speech Tagging* Bahasa Arab menggunakan teknik *HMM*

**ABSTRAK**

*Part-of-speech (POS) Tagger* memainkan peran penting dalam *Natural Language Applications* seperti *Speech Recognition*, *Natural Language Parsing*, *Information Retrieval* dan *Multi Words* *Term Extraction*. Penelitian ini mengusulkan pembangunan teknik *POS Tagging* yang efisien dan akurat untuk bahasa Arab menggunakan pendekatan statistik. Metode *Berbasis Aturan* Arab menderita kata-kata yang salah klasifikasi dan tidak dianalisis karena masalah *ambiguitas*. Untuk mengatasi dua masalah ini, kami mengusulkan *Hidden Markov Model (HMM)* yang terintegrasi dengan metode *Berbasis Aturan* Arab. *POS Tagger* kami menghasilkan satu set 4 tag POS: Noun, Verb, Particle, dan Quranic Initial (INL). Teknik yang diusulkan menggunakan informasi kontekstual yang berbeda dari kata-kata dengan berbagai fitur yang membantu untuk memprediksi berbagai kelas *POS*. Untuk mengevaluasi keakuratannya, metode yang diusulkan telah dilatih dan diuji dengan Holy Quran Corpus yang berisi 77 430 istilah untuk bahasa Arab Klasik yang tidak terakreditasi. Hasil eksperimen menunjukkan efisiensi metode kami untuk P*OS Tagging Arab*. Keakuratan yang diperoleh adalah 97,6% dan 94,4% untuk masing-masing metode kami dan untuk metode tagger berdasarkan Peraturan.

**KATA KUNCI**   
*Natural Language Applications*, *Natural Language Parsing*, *part-of-speech Tagger*, *Hidden Markov Model*, *Speech Recognition*.

**1. PERKENALAN**

*POS tagging*, juga disebut disambiguasi kategori kata atau penandaan gramatikal, adalah proses menganotasi kata-kata yang terjadi dalam teks dengan Part-Of-Spech yang terkait yang menangkap hubungan antara kata-kata ini dan kata-kata terkait yang berdekatan dalam kalimat, frasa, atau paragraf [1] [2] [3]. Penandaan Part Of Speech telah mendapat banyak arti penting di bidang Natural Language Processing (NLP) seperti ekstraksi informasi, penguraian teks dan pemrosesan semantik.

Arabic POS Tagging adalah proses identifikasi kategori leksikal kata Arab yang ada dalam sebuah kalimat berdasarkan konteksnya [5]. Kategori yang paling banyak digunakan adalah nomina, kata keterangan, kata kerja dan kata sifat. Ini dilakukan atas dasar peran kata-kata, baik secara individu maupun dalam kalimat. Sebagian besar kata yang muncul dalam teks Arab yang tidak terakreditasi memiliki ambiguitas dalam hal bagian pembicaraan mereka [4]. Ambil contoh istilah "ذھب", itu bisa diperlakukan sebagai kata benda "emas" atau kata kerja "pergi".

Ada tiga pendekatan umum untuk menangani masalah pemberian tag: Pendekatan berbasis aturan, pendekatan statistik, dan pendekatan Hybrid. Pendekatan berbasis aturan terdiri dari mengembangkan basis pengetahuan aturan yang ditetapkan oleh ahli bahasa untuk menentukan secara tepat bagaimana dan di mana untuk menetapkan berbagai tag POS. Pendekatan statistik terdiri dari membangun model yang dapat dilatih dan menggunakan corpus yang ditandai sebelumnya untuk memperkirakan parameternya. Setelah ini selesai, model dapat digunakan untuk menentukan tagger dari teks lain. Umumnya, penentu statistik yang sukses terutama didasarkan pada Hidden Markov Model (HMMs). Akhirnya, pendekatan hibrida terdiri dalam menggabungkan pendekatan berbasis aturan dengan pendekatan statistik. Baru-baru ini, sebagian besar Taggers POS menggunakan pendekatan terakhir karena memberikan hasil yang lebih baik.

Di antara karya-karya terbaru, kami lebih memilih metode berbasis aturan yang diusulkan oleh A.Taani [19] dibandingkan metode lain karena sejumlah alasan. Pertama, mudah dimengerti, akurat, dan bergantung pada struktur kalimat bahasa Arab yang benar menggunakan metrik pola sintaksis. Kedua, tidak seperti tagger lain yang umumnya dikembangkan untuk Bahasa Arab Standar Modern (MSA) dan dengan demikian mungkin tidak sesuai untuk Bahasa Arab Klasik (CA), metode Taani dapat berhubungan langsung dengan CA yang merupakan bahasa dari Al-Qur'an Suci.

Namun, metode berbasis aturan ini [19] menyajikan beberapa kelemahan: itu mungkin misclassify dan unanalyze beberapa kata. Misalnya, istilah "ھدى" (yaitu "atas, kanan") tidak dianalisis dan kata "موتكم" (yaitu "kematian Anda") diberikan oleh tag "kata kerja" yang salah.

Untuk mengatasi masalah ini, kami mengusulkan metode Penandaan Bagian-Of-Speech Arab berdasarkan pendekatan hibrida yang menggabungkan pendekatan berbasis aturan dengan pendekatan statistik yang bergantung pada struktur kalimat bahasa Arab yang meningkatkan Tagging Pos Taani [19]. Sisa kertas diatur sebagai berikut: Pada bagian 2, kami menggambarkan pekerjaan terkait teknik pemberian tag POS dalam bahasa Arab. Pemberi berdasarkan aturan disajikan dalam bagian 3. Bagian 4 menjelaskan prinsip-prinsip tag HMM. Metode yang kami usulkan diuraikan pada bagian 5. Bagian 6 menyajikan hasil eksperimen. Akhirnya, Bagian 7 menyimpulkan makalah dan menjelaskan pekerjaan di masa depan.

**2. KARYA TERKAIT**

Pemberian tag bagian-terdiri dari penugasan ke setiap kata dari sebuah kalimat tag yang menunjukkan fungsi kata dalam konteks tertentu. Seperti yang telah kami sebutkan sebelumnya, ada banyak metode penyihir tagging POS dapat diklasifikasikan dalam tiga kategori: Pendekatan statistik, pendekatan Rulebased dan pendekatan Hybrid.

**2.1. Pendekatan statistik:**

Pendekatan statistik membutuhkan usaha manusia yang jauh lebih sedikit, model yang berhasil selama tahun-tahun terakhir Model Markov Tersembunyi dan teknik-teknik terkait telah berfokus pada pembuatan model probabilistik dari urutan transisi tag dalam kalimat. Tugas ini sulit untuk bahasa Arab karena kurangnya corpus besar beranotasi. Sejauh ini, banyak metode pemberian tag PO telah disajikan dalam bahasa Arab yang seringkali bersifat statistik. Banko dkk. [16] menyajikan tag HMM yang mengeksploitasi konteks di kedua sisi kata yang akan diberi tag. Ini dievaluasi baik dalam kasus yang tidak diawasi dan diawasi. Tagger Orumchian [10] disajikan untuk penandaan POS Persia yang mengikuti tag tag TNT POS. Penanda TNT didasarkan pada teori Model Markov Tersembunyi. Sistem ini menggunakan 2,5 juta kata tag sebagai data pelatihan dan ukuran set tag adalah 38. Guessoum et al [26] menyajikan sistem penandaan POS untuk menyelesaikan ambiguitas melalui penggunaan statistik.

model bahasa dikembangkan dari corpus Arab sebagai Hidden Markov Model (HMM). Albared dkk. [27] mengembangkan Model Markov Hidden Markov (HMM) untuk mengatasi masalah penandaan POS bahasa Arab. Parameter HMM diperkirakan dari data pelatihan yang kecil. Mereka telah mempelajari perilaku HMM yang diterapkan pada pemberian tag Pos Arab menggunakan sejumlah kecil data. Dengan menggunakan algoritma smoothing yang berbeda dengan model HMM untuk mengatasi masalah sparseness data. Alhadj et al. [30] mengusulkan metode baru pemberian tag bagian-dari-pidato yang dapat digunakan untuk menganalisis dan menambahkan catatan teks-teks Arab tradisional, terutama teks Al-Qur'an. Pendekatan ini menggabungkan analisis morfologis dengan Model Markov Tersembunyi (HMM) berdasarkan struktur kalimat bahasa Arab.

**2.2. Pendekatan berbasis aturan:**

Pendekatan ini telah berhasil digunakan dalam mengembangkan banyak sistem pemrosesan bahasa alami. Sistem yang menggunakan transformasi berbasis aturan didasarkan pada inti pengetahuan linguistik yang solid. Salah satunya adalah imbuhan. Beberapa imbuhan tepat untuk kata kerja; beberapa layak untuk kata benda; dan beberapa lainnya digunakan dengan kata kerja dan kata benda. Tanda lain yang penting dalam bahasa Arab adalah polanya, yang merupakan panduan penting dalam mengenali kategori kata. Pendekatan ini juga digunakan untuk beberapa tugas tertentu. Diab dkk. [15] merancang sistem penandaan otomatis untuk mengesahkan sebagian dari tag khusus dalam teks Arab. Habash dkk. [24] mengusulkan penganalisis morfologis untuk mengesampingkan dan menandai kata-kata Arab secara morfologis. Freeman [13] mendeskripsikan sistem pemberian tag bagian-dari-bahasa Arab berdasarkan sistem penandaan Brill yang merupakan sistem pembelajaran mesin yang dapat dilatih dengan korpus yang ditandai sebelumnya. Penulis menggunakan set tag yang berisi 146 tag yang diekstraksi dari Brown corpus untuk bahasa Inggris. Lee et al. [18] menggunakan korpus kata-kata tersegmentasi secara manual yang tampaknya merupakan bagian dari rilis pertama ATB (110.000 kata). Mereka memperoleh daftar prefiks dan sufiks dari corpus ini yang tampaknya ditambah dengan daftar afiks yang diturunkan secara manual. Maamouri dkk. [14] menyajikan sistem pemberian tag bagian-dari-bicara untuk bahasa Arab. Para penulis mendasarkan pekerjaan mereka pada hasil penganalisis morfologi Tim Buckwalter. Sistem penandaan ini diuji pada korpus yang terdiri dari 734 file yang diekstrak dari "Agence France Press" [25].

**2.3. Pendekatan hibrida:**

Pendekatan ini terdiri untuk menggabungkan metode berbasis aturan dan metode statistik yang digunakan untuk menetapkan tag terbaik untuk setiap kata-kata input teks. Untuk metode hibrida, para penija Arab yang berbeda baru saja muncul. Di antara studi-studi ini, Khoja [12] menggabungkan teknik statistik dan teknik berbasis aturan dan menggunakan set tag dari 131 yang pada dasarnya berasal dari set tag bahasa Inggris BNC. TliliGuiassa [17] menggunakan metode hibrida berdasarkan aturan dan metode pembelajaran berbasis memori. Salah satu karya terbaru untuk penandaan POS dilakukan oleh Jabbari dan Allison [7]. Pendekatan mereka adalah berdasarkan transformasi dan sebelumnya telah digunakan dalam bahasa Inggris oleh Brill dan Hepple [8, 9]. Konstruksi tag ini berisi mesin pembelajar terlatih yang mencakup aturan yang diperkirakan. Bahkan, mereka menerapkan implementasi Error-Driven Transformation Based Learning.

Perhatikan bahwa sebagian besar dari taggers ini [14] menerapkan terjemahan teks asli Arab ke bahasa Inggris dan menggunakan set tag yang berasal dari bahasa Inggris yang tidak sesuai untuk bahasa Arab. Tagger lain [12] [13] [16] bergantung pada transliterasi teks masukan bahasa Arab. Selain itu, sebagian besar tagger umumnya dikembangkan untuk Bahasa Arab Standar Modern (MSA) dan dengan demikian mungkin tidak sesuai untuk Bahasa Arab Klasik yang merupakan bahasa Al-Qur'an. Umumnya, Metode hibrida [7] [8] [9] memberikan hasil yang lebih baik untuk pemberian tag POS.

Di antara karya-karya terbaru, kami telah memilih metode berbasis aturan yang disajikan oleh A.Taani [19] karena akurat dan bergantung pada struktur kalimat bahasa Arab yang benar menggunakan metrik pola sintaksis. Selain itu, ia menggunakan Bahasa Arab Klasik (CA) yang merupakan bahasa Al-Qur'an.

Namun, metode berbasis aturan ini [19] menyajikan beberapa masalah: mungkin salah mengartikan dan tidak menganalisis beberapa kata. Misalnya, istilah "ھدى" (yaitu "atas" atau "kanan") tidak dianalisis dan kata "موتكم" (yaitu "kematian Anda") diberikan oleh tag "kata kerja" yang salah. Untuk mengatasi masalah ini, makalah ini mengusulkan metode Penandaan Bagian-Of-Speech Arab yang menggabungkan pendekatan Berbasis Peraturan [19] dengan teknik HMM. Yang terakhir ini lebih unggul dari model lain dalam hal waktu pelatihan dan cocok untuk aplikasi yang berhubungan dengan sejumlah besar teks. Dalam dua bagian berikut, baik Metode Penandaan Berbasis Peraturan Taani dan HMM akan disajikan dan dirinci.

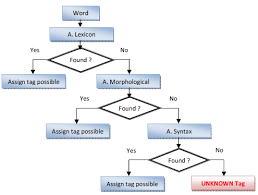
**3. METODE TAGGING ATURAN ATURAN TAANI**

Metode penandaan berbasis aturan [19] memungkinkan penggolongan kata-kata dalam teks Arab non-suara ke tag mereka. Ini terdiri dari tiga fase utama: analisa leksikon, penganalisis morfologi, dan penganalisis sintaks. Gambar 1 menunjukkan arsitektur sistem ini.

Lexicon Analyzer: Pada langkah ini, leksikon daftar berhenti dalam bahasa Arab didefinisikan. Leksikon ini termasuk kata depan, kata keterangan, konjungsi, partikel interogatif, pengecualian dan interjections. Semua kata harus melewati fase ini. Jika kata tersebut ditemukan dalam leksikon, itu dianggap sebagai tag. Lain, itu lolos ke langkah berikutnya.

Penganalisis morfologi: Setiap kata yang belum ditandai pada fase sebelumnya akan berimigrasi ke fase ini. Satu set afiks dari setiap kata diekstraksi. Afiks dapat berupa awalan, akhiran, atau infiks. Setelah itu, imbuhan dan hubungan di antara mereka digunakan dalam satu set aturan untuk menandai kata itu ke dalam kelasnya. Bukannya fase ini adalah inti dari sistem, karena ini membedakan persentase utama kata-kata yang tidak terikat ke dalam kata benda atau kata kerja.

Syntax analyzer: Fase ini dapat membantu dalam menandai kata-kata yang gagal di-tag dua fase sebelumnya. Ini terdiri dari dua aturan: konteks kalimat dan parsing balik. Aturan konteks kalimat didasarkan pada hubungan antara kata-kata tanpa tag dan yang berdekatan. Bahasa Arab memiliki beberapa jenis hubungan antara kata-kata yang berdekatan. Misalnya preposisi dan interjections selalu diikuti oleh kata benda. Hubungan ini memungkinkan untuk menandai kata-kata ke dalam kelas terkait. Aturan parsing terbalik didasarkan pada tata bahasa bebas konteks Arab. Para penulis mengusulkan seperangkat aturan yang sering digunakan dalam bahasa Arab.



Gambar 1. Arsitektur rule-Based Arabic POS Tagger [19]

Pada bagian berikut, kami menyajikan model HMM karena akan diintegrasikan dalam metode kami untuk teks tagging Arab POS.

**4. MODEL MARKOV TERSEMBUNYI**

Penggunaan *Hidden Markov Model (HMM)* untuk melakukan *POS Tagging* dapat dilihat sebagai kasus khusus inferensi Bayesian [20]. Ini dapat diformalkan sebagai berikut: untuk urutan kata tertentu, apa urutan tag terbaik yang sesuai dengan urutan kata-kata ini? Jika kami merepresentasikan teks yang dimasukkan (urutan unit morfologi dalam kasus kami) oleh W = (wi)1≤i≤n dan urutan tag dari leksikon dengan T = (ti) 1≤i≤n , kita harus menghitung:

(1)

Dengan menggunakan aturan Bayesian dan kemudian menghilangkan bagian konstan, persamaan dapat ditransformasikan ke yang baru ini:

(2)

Di mana mewakili probabilitas dari urutan tag (probabilitas transisi tag), dan dapat dihitung menggunakan model N-gram, sebagai berikut:

(3)

Corpus pelatihan yang ditandai digunakan untuk menghitung P(ti | ti-n ... ti-2 ti-1), dengan menghitung frekuensi N-gram sebagai berikut:

(4)

Namun, dapat terjadi bahwa beberapa trigram (atau bigrams) tidak akan pernah muncul di set pelatihan; jadi, untuk menghindari menugaskan probabilitas nol ke trigram tak terlihat (bigrams), kami menggunakan interpolasi yang dihapus yang dikembangkan oleh [20]:

(5)  
Dimana λ1 + λ2 + ... + λn = 1

Kemudian, untuk menghitung kemungkinan dari kata urutan yang diberikan tag, probabilitas dari suatu kata yang muncul umumnya seharusnya hanya bergantung pada POS Tag nya sendiri. Jadi, itu bisa ditulis sebagai berikut:

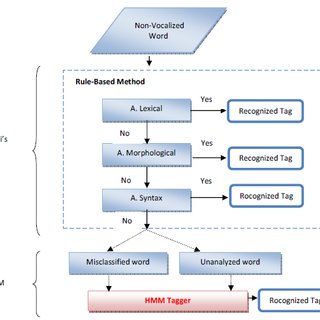
(6)  
Selain itu, satu set pelatihan yang ditandai harus digunakan untuk menghitung probabilitas ini, sebagai berikut:

(7)  
Probabilitas urutan tag dan kata likelihood mewakili parameter model HMM: probabilitas transisi dan emisi (observasi) probabilitas. Setelah parameter-parameter ini ditetapkan, model HMM dapat digunakan untuk menemukan urutan terbaik dari suatu urutan kata-kata masukan. Algoritma Viterbi dapat digunakan untuk melakukan tugas ini.

**5. METODE YANG DIUSULKAN UNTUK POS TAGGING ARAB**

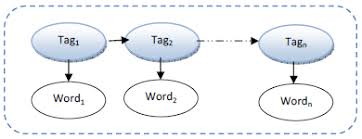
Metode yang diusulkan didasarkan pada pendekatan hibrida; ini menggabungkan metode Rule-Based yang disajikan oleh Taani [19] dengan model HMM (lihat Gambar 2). Seperti yang telah kami sebutkan, metode berbasis Kaidah disusun oleh tiga langkah: analisa leksikon, penganalisis morfologi dan penganalisis sintaksis (lihat bagian 3).

Hampir semua kata diakui oleh metode berbasis aturan. Namun, beberapa istilah tidak dianalisis atau salah diklasifikasi. Istilah-istilah ini (istilah sisanya gagal) akan dianalisis menggunakan model HMM. Keadaan model ini sesuai dengan tag bagian-dari-pidato dan pengamatan sesuai dengan kata-kata (lihat Gambar 3). Ide dasar dari model HMM kami yang mengadopsi pembelajaran yang diawasi adalah menetapkan tag yang paling mungkin ke kata dari kalimat input. Diperlukan dua langkah utama: langkah pelatihan dan langkah uji.



Gambar 2. Arsitektur metode kami POS Tagger Bahasa Arab

Langkah pelatihan didasarkan pada pembelajaran yang diawasi. Hal ini memungkinkan untuk mempelajari parameter dari model HMM menggunakan korpus dengan memperkirakan transisi dan probabilitas emisi. Pertama, untuk setiap iterasi (mengenai satu istilah) kita menghitung probabilitas emisi untuk setiap tag yaitu p (word | tagi) (lihat persamaan 7). Kedua, untuk setiap iterasi (mengenai satu tag) kami menghitung probabilitas transisi yang mewakili hubungan antara tag dan tag sebelumnya yaitu p (tag | tag sebelumnya) (lihat persamaan 7) untuk Model Markov Tersembunyi. Hasil dari langkah ini adalah dua matriks: matriks probabilitas transisi (Tag / Tag) dan matriks probabilitas emisi (kata / Tag).



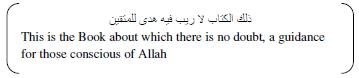
Gambar 3: Arsitektur HMM Tagger

Langkah pengujian bertujuan untuk menetapkan tag kata kemungkinan terbaik yang telah mengalami kesalahan klasifikasi atau tidak dianalisis selama proses berbasis aturan. Pertama, kami memberikan semua tag yang mungkin untuk kata ini. Kemudian, kami menghitung probabilitas masing-masing tag kata ini dengan menggunakan probabilitas transisi dan probabilitas emisi. Akhirnya, algoritma Viterbi digunakan untuk menghitung jalur kemungkinan terbaik (urutan tag terbaik) untuk kata tertentu dalam urutan (kalimat).

Untuk mengilustrasikan langkah-langkah ini, kami mempertimbangkan sebagai contoh kalimat berikut:

**ذَٰلِكَ الْكِتَابُ لَا رَيْبَ ۛ فِيهِ ۛ هُدًى لِّلْمُتَّقِينَ**

Kitab (Al Quran) ini tidak ada keraguan padanya, petunjuk bagi mereka yang bertakwa,

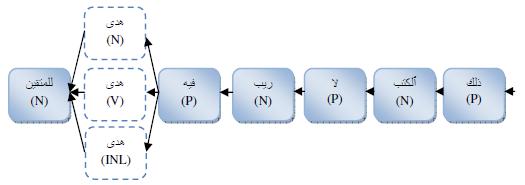


Gambar 4: Contoh kalimat termasuk kata Arab dengan tag yang tidak diketahui

Untuk kata "ھدى", tagnya tidak dikenal dengan metode berbasis aturan. Pertama, kami menugaskannya ke tiga label berbeda N, V, dan INL. Dengan menggunakan model HMM, transisi dan probabilitas emisi yang sebelumnya diperkirakan dalam fase pelatihan, kami menghitung probabilitas masing-masing tag untuk kata "ھدى" sebagai fungsi probabilitas tag sebelumnya dalam kalimat:  
   
- P (ھدى \ N) = P (ھدى \ N) P (N \ P) P (P \ N) P (N \ P) P (P \ N) P (N \ P)

- P (ھدى \ V ) = (ھدى \ V) P (N \ P) P (P \ N) P (N \ P) P (P \ N) P (N \ P)

- P (ھدى \ INL) = (ھدى \ INL) P (N \ P) P (P \ N) P (N \ P) P (P \ N) P (N \ P)



Akhirnya, kami menghitung jalur kemungkinan terbaik (yaitu urutan tag terbaik) dengan menggunakan algoritma Viterbi. Yang terakhir adalah algoritma decoding paling umum untuk HMM yang memberikan urutan tag yang paling mungkin diberikan satu set tag. Ini menggunakan rumus berikut.

**6. HASIL DAN PEMBAHASAN**

Pada bagian ini, kami menyajikan dan menggambarkan corpora yang digunakan untuk mengevaluasi metode yang kami usulkan untuk Tagging POS Arab. Kemudian, kami menyajikan beberapa tugas pra-pemrosesan yang dilakukan pada korpus, dan mendeskripsikan set tag yang kami gunakan. Hasil eksperimen akan disajikan dan dibahas.

**6.1. Deskripsi Corpus**

Kami menggabungkan Quran Arabic Corpus [22] yang dinamakan sebagai: "quranic-corpus-text-0.2". Bentuk bahasa korpus adalah bahasa Arab klasik (CA). Alfabet bahasa Arab dari korpus terdiri dari huruf-huruf berikut: أ ب ت ث د ج ح خ ه ع غ ف ق ص ض ط ك م ن ل ي س ش ظ ز و ة ى لا ر ؤ ء ئ. Transkripsi fonetik untuk huruf ditunjukkan pada Tabel 1.

The Holy Quran corpus terdiri dari 6236 kalimat dengan total 77430 kata dan menggunakan 33-tag [23]. Set tag ini terdiri dari: Noun (N), Proper Noun (PN), Number(NUM), Adjective (ADJ), Imperative verbal noun (IMPN), Verb (V), Prohibition Particle (PRO), Negative Particle (NEG), Accusative Particle (ACC), Conditional Particle (COND), Restriction Particle (RES), Particle of Certainty (CERT), Interrogative Particle (INTG), Inceptive Particle (INC), Vocative Particle (VOC), Retraction Particle (RET), Amendment Particle (AMD), Future Particle (FUT), Exhortation Particle (EXH), Exceptive Particle (EXP), Explanation Particle (EXL), Surprise Particle (SUR), Aversion Particle (AVR), Answer Particle (ANS), Coordinating Conjunction (CONJ), Subordinating Conjunction (SUB), Time Adverb (T), Location Adverb (LOC), Personal Pronoun (PRON), Relative Pronoun (REL), Demonstrative Pronoun (DEM), Quranic Initial ( INL).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Huruf** | **Arab** | **Tulisan** |  | **Huruf** | **Arab** | **Tulisan** |  |
|  |  |
| alif | ا | A |  | za | ظ | Z |  |
| ba | ب | B |  | ain | ع | ‘ |  |
| ta | ت | T |  | Ghayn | غ | Gh |  |
| Tha | ث | Th |  | Fa | ف | F |  |
| Jim | ج | J |  | Qaf | ق | Q |  |
| Ha | ح | h |  | Kaf | ك | K |  |
| Kha | خ | Kh |  | Lam | ل | L |  |
| Dal | د | D |  | Mim | م | M |  |
| Dhal | د | dh |  | Nun | ن | N |  |
| Ra | ر | R |  | Ha | ه | H |  |
| Zay | ز | Z |  | Waw | و | W |  |
| Sin | س | S |  | Ya | ي | Y |  |
| Shin | ش | Sh |  | Hamza | ء | ‘ |  |
| Sad | ص | s |  | Alif maksura | ى | A |  |
| Dad | ض | d |  | Ta marbuta | ة | T |  |
| Ta | ط | t |  |  |  |  |  |

Bentuk korpus yang tidak terakreditasi disimpan dalam file baru dengan menghapus diakritik untuk semua kata-kata korpus. Penting untuk dicatat bahwa bentuk tidak terakreditasi dari corpus Holy Quran digunakan dalam beberapa studi NLP. Hasil eksperimen pada bahasa Arab yang tidak terakreditasi berguna karena tulisan Arab kebanyakan ditulis tanpa diakritik

|  |  |
| --- | --- |
| **Tag Set Komprehensif (33 tag)** | **Tag Set disederhanakan**  **(4 tag)** |
| N, PN, NUM, ADJ, IMPN | N |
| V | V |
| PRO, NEG, ACC, COND, RES, CERT, INTG, INC, VOC, RET, AMD,  FUT, EXH, EXP, EXL, SUR, AVR, ANS, CONJ, SUB, T, LOC,  Pron, REL, DEM | P |
| INL | INL |

Pekerjaan lain juga dilakukan ke alat NLTK [21] untuk memproses file korpus Al-Qur'an menggunakan set tag yang disederhanakan. Set tag yang disederhanakan hanya mencakup 4-tag yaitu: Noun (N), Verb (V) Particle (P) dan Quranic Initial (INL). Tabel 2 menyajikan kriteria pemetaan yang digunakan untuk mengkonversi set tag komprehensif (33 tag) dari korpus asli ke yang disederhanakan (4 tag).

Kami menerapkan modul Python baru untuk mengintegrasikan file yang dibuat baru dari corpus Holy Quran ke dalam alat NLTK untuk melakukan eksperimen kami.

**6.2. Kumpulan data dan evaluasi**

Kami melakukan beberapa eksperimen menggunakan corpus sebelumnya. Percobaan didasarkan pada bahasa Arab klasik untuk bentuk yang tidak terakreditasi sesuai dengan set 4-tag. Tabel3 menyajikan beberapa hasil dari tagger POS Arab menggunakan dua Taggers: Taani’s Rule-based dan metode yang kami usulkan.

Dalam tabel ini untuk setiap kalimat kita hanya mempertimbangkan istilah yang ambigu (kata-kata yang salah diklasifikasikan dan tidak dianalisis). Misalnya kalimat yang dibentuk oleh 15 kata:

(3:59). ”إن مثل عيسى عند ﷲ كمثل آدم خلقه من تراب ثم قال له كن فيكون”

”Sesungguhnya perumpamaan (penciptaan) Isa bagi Allah, seperti (penciptaan) Adam. Dia menciptakannya dari tanah, kemudian Dia berkata kepadanya, “Jadilah!” Maka jadilah sesuatu itu.

berisi satu kata (Jadilah) "كن" yang salah diklasifikasi oleh metode Taani dan dua istilah (Isa) "عيسى" dan (Adam) "ءادم" yang tidak diselesaikan dengan metode Taani. Namun istilah-istilah ini diperlakukan dengan benar oleh metode kami.

Tabel 3. Contoh hasil yang diperoleh menggunakan tagger Berbasis Aturan dan tagger Hybrid kami

|  |  |  |  |
| --- | --- | --- | --- |
| Kalimat | Istilah | Metode Taani | Metode yang diusulkan |
| V/ خلقه ?/ آدم N/ كمثل N/ ﷲ P/ عند ?/ عيسى N/ مثل P/ إن  V/ فيكون ?/ كن P/ قال V/ ثم P/اب N/تر P/ من | عيسى | ? | N |
| ءادم | ? | N |
| كن | N | V |
| N/ للمتقين ?/ ھدى P/ فيه N/ ريب P/ لا N/ ٱلكتب P/ ذلك | ھدى | ? | N |
| N/وبﭑليوم N/ باللّه ?/ أمنا V/ يقو P/ من N/ الناس P/ ومن  ?/ بمؤمنين P/هم P/وما N/ ٱلءاخر | ءامنا | ? | V |
| بمؤمنين | V | N |
| V /ويسفك P/ فيھا V /يفسد P/ من P/ فيھا ?/ أتجعل V / قالوا  P/ لك V /ونقدس N /بحمدك ?/ نسبح P/ ونحن N/ ٱلدماء | أتجعل | ? | V |
| نسبح | N | V |
| V /يسومونكم N / فرعون ءال/ ? P/ من V /نجينكم P/ وإذ  N /نساءكم ?/ ويستحيون N/ أبناءكم ?/ يذبحون N /ٱلعذاب N /ﺳوء  N/عظيم N /ربكم P/ من N /بلاء P/ ذلكم P/ وفى | ءال | ? | N |
| يذبحون | N | V |
| يستحيون | N | V |
| ?/ تشكرون P/ لعلكم ?/ موتكم N /بعد P/ من V /بعثنكم P/ ﺛم | موتكم | V | N |
| تشكرون | N | V |

Tabel 4 menyajikan akurasi yang diperoleh dari metode Rule-Based dari Taani dan metode yang kami usulkan, dengan nilai yang berbeda dari persentase training corpus. Dari tabel ini, kami menunjukkan bahwa tager POS yang kami usulkan melebihi Tagger berbasis aturan dalam hal akurasi. Ukuran data pelatihan meningkat dengan variasi corpus pelatihan yang berbeda. Hasil yang diperoleh untuk metode kami mencapai kinerja yang lebih baik: 97% vs 94% untuk metode Taani.

|  |  |  |
| --- | --- | --- |
| **metode** | **Metode Taani ini** | **metode kami** |
| **pelatihan%** |  |  |
| **30%** | 94% | 97% |
| **70%** | 94,40% | 97,40% |
| **80%** | 94,20% | 97,60% |
| **90%** | 94,40% | 97,60% |

Saat kami melatih tagger pada sejumlah besar data, kami mendapatkan hasil pemberian tag yang akurat. Jadi kami menyimpulkan bahwa hasil bergantung pada sebagian kecil data pelatihan yang digunakan untuk melatih Tagger. Oleh karena itu mempertimbangkan ukuran korpus yang digunakan untuk percobaan, tagger kami mencapai akurasi yang luar biasa dengan corpus Al-Quran Suci dibandingkan dengan metode Taani.

**7. KESIMPULAN DAN PEKERJAAN DI MASA DEPAN**

Dalam makalah ini, kami telah menyajikan teknik Tagging part-of-speech (POS) yang efisien dan akurat untuk bahasa Arab menggunakan pendekatan statistik. Pembuat tag yang dikembangkan menggunakan pendekatan yang menggabungkan metode berbasis Peraturan Taani dengan Model Tersembunyi Markov (HMM). Sebuah arsitektur yang sesuai dari model HMM ditentukan berdasarkan struktur kalimat yang memungkinkan kita untuk menangani dengan benar ambiguitas yang terkait dengan kata-kata yang salah klasifikasi dan tidak dianalisis dalam metode RuleBased Arab. Untuk mengevaluasi keakuratan yang diusulkan POS Tagger, serangkaian percobaan dilakukan menggunakan Holy Quran corpus yang mengandung 77 430 istilah untuk bahasa Arab Klasik yang tidak terakreditasi.

Percobaan dilakukan dengan melakukan tes lebih lanjut pada dataset yang lebih menarik untuk mengevaluasi kinerja nyata dari pendekatan ini. Akurasi sekitar 97,6% menunjukkan hasil yang sangat baik dari metode kami dibandingkan dengan Aturan Berbasis Taani. Kami mencatat bahwa keakuratan sedikit meningkat dengan meningkatnya jumlah kata dalam korpus pelatihan. Di masa depan, kami berencana untuk meningkatkan ketepatan penandaan kata-kata yang tidak dikenal dengan menggunakan korpus pelatihan lain, dan menerapkan tagger POS kami dalam ekstraksi Multi-Word Terms.

REFERENCES

1. Lee, S.hyun. & Kim Mi Na, (2008) “This is my paper”, ABC Transactions on ECE, Vol. 10, No. 5, pp120-122.
2. Gizem, Aksahya & Ayese, Ozcan (2009) Coomunications & Networks, Network Books, ABC Publishers.
3. <http://en.wikipedia.org/wiki/Part-of-speech_tagging>.
4. L.Van Guilder, (1995) “Automated Part of Speech Tagging: A Brief Overview” Handout for LING361, Georgetown University.
5. H. Halteren, J.Zavrel & Walter Daelemans (2001).Improving Accuracy in NLP Through Combination of Machine Learning Systems. Computational Linguistics. 27(2): 199–229.
6. DeRose & J.Steven (1990) "Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages." PhD.Dissertation. Providence, RI: Brown University Department of Cognitive and Linguistic Sciences.
7. N. kumar Kumar, Anikel Dalal &Uma Sawant (2006)”hindi part of speech tagging and chunking”, NLPAI machine learning contest.
8. M. Mohseni, H. Motalebi, B. Minaei-bidgoli & M. Shokrollahi-far (2008) “A farsi part-of-speech tagger based on markov”. In the proceedings of ACM symposium on Applied computing, Brazil.
9. S. Jabbari & B. Allison (2007) “Persian Part of Speech Tagging”, In the Proceedings of Workshop on Computational Approaches to Arabic Script-Based Languages (CAASL-2), USA.
10. E. Brill (1995) “Transformation-Based Error-Driven Learning and Natural Language Processing: A case Study in Part of Speech Tagging”, Computational Linguistics, USA.
11. M. Hepple (2000), ”Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based Part of-Speech Taggers”, In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL). Hong Kong.
12. T. Brants (200),“TNT – a Statistical Part-of-Speech Tagger”, In the Proceedings of 6th conference on applied natural language processing (ANLP), USA.
13. K. Megerdoomian (2004), “Developing a Persian part-of speech tagger”, In the Proceedings of first Workshop on Persian Language and computer, Iran .
14. Khoja, S.( 2001) “ APT: Arabic part-of-speech tagger”. Proceeding of the Student Workshop at the 2nd Meeting of the NAACL, (NAACL’01), Carnegie Mellon University, Pennsylvania, pp: 1-6. <http://zeus.cs.pacificu.edu/shereen/NAACL.pdf>
15. Freeman A (2001), “Brill’s POS tagger and a morphology parser for Arabic”, In ACL’01 Workshop on Arabic language processing.
16. Maamouri M, Cieri C. (2002). “Resources for Arabic Natural Language Processing at the LDC”, Proceedings of the International Symposium on the Processing of Arabic,Tunisia, pp.125-146.
17. Diab M., Hacioglu K. and Jurafsky D. (2004), “Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks”. proc. of HLTNAACL’04: 149–152.
18. Banko M, Moore R. C. (2004). “Part of Speech Tagging in Context”, Proc of the 20th international conference on Computational Linguistics, Switzerland.
19. Tlili-Guiassa Y. (2006) “Hybrid Method for Tagging Arabic Text”. Journal of Computer Science 2 (3): 245-248.
20. L. Young-Suk, K. Papineni & S. Roukos ( 2003), “Language Model Based Arabic Word Segmentation,” in Proceedings of the Annual Meeting on Association for Computational Linguistics, Japan, pp. 399- 406.
21. A.T Al-Taani & S. Abu-Al-Rub (2009),”A rule-based approaches for tagging non-vocalized Arabic words”. The International Arab Journal of Information Technology, Volume6 (3): 320-328.
22. T. Brants (2000),” TnT: A statistical part of speech tagger”, Proceedings of the 6th Conference on Applied Natural Language Processing, Apr. 29- May 04, Association for Computational Linguistics Morristown, New Jersey, USA., pp: 224-231.
23. NLTK, Natural Language Toolkit. <http://www.nltk.org/Home>
24. Quranic Arabic Corpus: <http://corpus.quran.com>
25. Quran Tagset: http://corpus.quran.com/documentation/tagset.jsp
26. N. Habash & O. Rambow (2005), “Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop,” in Proceedings of the Annual Meeting on Association for Computational Linguistics, Michigan, pp. 573-580.
27. <http://sibawayh.emi.ac.ma/web/s/?q=node/79>
28. F. Al Shamsi & A.Guessoum(2006),” A Hidden Markov Model–Based POS Tagger for Arabic”, 8es Journées internationales d’Analyse statistique des Données Textuelles (JADT).
29. M. Albared & O.Nazlia(2010),” Automatic Part of Speech Tagging for Arabic: An Experiment Using Bigram Hidden Markov Model “,Springer-Verlag Berlin Heidelberg, LNAI 6401, pp. 361–370.
30. Y.O. Mohamed Elhadj(2009),” Statistical Part-of-Speech Tagger for Traditional Arabic Texts”, Journal of Computer Science 5 (11): 794-800