Pendekatan Berbasis Aturan untuk Menandai Kata-Kata Arab yang Tidak Dinvokalisasi

Ahmad Al-Taani and Salah Abu Al-Rub Department of Computer Sciences, Yarmouk University, Jordan

Abstrak:

Dalam karya ini, kami menyajikan sistem penandaan yang mengklasifikasikan kata-kata dalam teks Arab yang tidak disuarakan ke tag mereka. Sistem penandaan yang diusulkan melewati tiga tingkat analisis. Tingkat pertama adalah penganalisa leksikal yang terdiri dari leksikon yang mengandung semua kata dan partikel tetap seperti preposisi dan kata ganti. Tingkat kedua adalah penganalisa morfologis yang bergantung pada struktur kata menggunakan pola dan imbuhan untuk menentukan kelas kata. Tingkat ketiga adalah penganalisis sintaksis atau penandaan gramatikal yang bergantung pada proses pemberian tag gramatikal pada kata-kata berdasarkan konteksnya atau posisi kata dalam kalimat. Level analisa sintaksis terdiri dari dua tahap: tahap pertama tergantung pada kata kunci spesifik yang menginformasikan tag kata yang berurutan, tahap kedua adalah teknik penguraian terbalik yang memindai tata bahasa Arab yang tersedia untuk mendapatkan kelas kata ambiguitas tunggal. dalam kalimat. Kami telah menguji sistem yang diusulkan pada corpus terdiri dari 2355 kata. Hasil eksperimen menunjukkan bahwa sistem yang diusulkan mencapai tingkat keberhasilan mendekati 94% dari total jumlah kata dalam sampel yang digunakan dalam penelitian.

Kata kunci: Penandaan sebagian ucapan, penganalisis leksikal, penganalisa morfologis, pemrosesan bahasa Arab

1. Pendahuluan

Bahasa Arab menempati urutan keenam di liga bahasa dunia dengan perkiraan 200 juta penutur asli dan banyak digunakan di seluruh dunia Muslim [8]. Beberapa bukti arkeologis menunjukkan bahwa bahasa Arab mungkin merupakan bahasa tertua [27].

Teks-teks Arab dapat berupa teks yang disuarakan seperti bahasa Al-Quran atau teks yang tidak disuarakan yang digunakan di koran, buku, dan media. Menangani teks yang tidak disuarakan membingungkan karena kata yang tidak disuarakan mungkin memiliki lebih dari satu makna. misalnya, kata arab yang tidak disuarakan (كتب) memiliki tiga kemungkinan interpretasi: "kataba" (dia menulis), "kutiba" (telah ditulis), dan "kutubun" (buku).

Untuk memahami apa itu kelas kata, kita harus memahami ide untuk memasukkan hal-hal yang sama dalam kelompok atau kategori. Kami biasanya menggunakan tiga kategori untuk mengklasifikasikan semua kata yang digunakan dalam bahasa Arab: Kata benda, kata kerja, dan partikel [17]. Klasifikasi ini tidak sempurna dalam teks Arab yang tidak disuarakan. Terkadang, sulit untuk menentukan kategori mana yang dimiliki kata. Selain itu, kata yang sama mungkin termasuk kategori yang berbeda tergantung pada bagaimana kata itu digunakan.

Klasifikasi kata adalah proses pemberian tag pada kata-kata dan seringkali hanya satu langkah dalam aplikasi pemrosesan teks maka teks yang ditandai dapat digunakan untuk analisis yang lebih dalam. Corpus yang ditandai lebih bermanfaat daripada korpus yang tidak ditandai karena ada lebih banyak informasi daripada dalam teks mentah saja. Setelah corpus ditandai, ia dapat digunakan untuk mengekstrak informasi dari corpus. Ini kemudian dapat digunakan untuk membuat kamus dan tata bahasa suatu menggunakan data bahasa nyata. Tagged corpora juga berguna untuk analisis kuantitatif terperinci teks dan merupakan persiapan untuk tugas pemahaman bahasa alami tingkat tinggi seperti penguraian, semantik, dan terjemahan [20]. Parser harus tahu tag untuk setiap kata. Sebagian besar pendekatan sebelumnya menggunakan penandaan manual tetapi memiliki penandaan otomatis akan meningkatkan efisiensi dan kinerja parser. Informasi tentang kategori kata sangat membantu dalam memahami makna kata tersebut secara penuh dan mengetahui cara menggunakannya. Misalnya, sistem terjemahan mesin memproses teks input secara bertahap: de-format, analisis morfologis, disambiguasi penandaan klasifikasi kata, transfer struktural dangkal, transfer leksikal, generasi morfologis, dan format ulang. Klasifikasi kata adalah tahap dasar yang diperlukan dalam sistem terjemahan mesin apa pun. Memiliki sistem penandaan otomatis akan meningkatkan efisiensi dan kinerja sistem terjemahan.

2. Pernyataan Masalah

Beberapa masalah menggunakan afiks dalam klasifikasi kata ditemui oleh beberapa peneliti. Beberapa huruf yang tampak seperti imbuhan sebenarnya adalah bagian dari kata seperti dalam kata ( التقي ELTAGA -memenuhi) kami memperlakukan dua huruf pertama (ال) sebagai artikel dan kata tersebut diklasifikasikan sebagai kata benda, tetapi pada kenyataannya beberapa dari surat-surat ini adalah bagian dari kata dan kata tersebut adalah kata kerja [19]. Beberapa huruf (misalnya vokal panjang) dapat berubah menjadi huruf lain ketika afiks ditambahkan dan oleh karena itu huruf harus diubah kembali ketika afiks itu dihapus.

Beberapa kata dalam teks yang tidak disuarakan mungkin memiliki lebih dari satu tag (kata-kata yang mendua dan kata-kata yang tidak dikenal) di mana klasifikasi dapat bergantung pada makna kata. Ambiguitas bahasa Arab terletak pada level yang berbeda. Kami memberikan beberapa contoh kemungkinan kombinasi kategori tata bahasa dari kata-kata yang tidak diucapkan

Banyak kata kerja memiliki bentuk yang sama dengan kata benda (terutama di akar yang tidak memiliki imbuhan). Sebagai contoh, kata (ذهب) dapat berarti pergi dan diklasifikasikan sebagai kata kerja masa lalu atau mungkin berarti emas dan diklasifikasikan sebagai kata benda [7].

Pola kata lain yang mencakup kata benda dan kata sifat adalah pola partisipatif aktif dan pasif (مفعول, فاعل) dan turunannya. kasus-kasus ini kadang-kadang bahkan lebih rumit karena mereka juga dapat diklasifikasikan dari waktu ke waktu sebagai preposisi seperti dalam kata (داحل dalam) dan sebagai partisip dengan fungsi kata kerja seperti dalam kalimat (هوداخل) “dia akan masuk” [18].

Banyak kata kerja memiliki bentuk yang sama dengan kata sifat. Seringkali kata kerja non-vokal dengan tiga radikal memiliki pola yang sama dengan kata sifat. Misalnya, tiga radikal “فرح” keduanya bisa berdiri untuk kata kerja “فٌرحَ” dan kata sifat “فٌرَخْ” [18]

Perpaduan yang paling penting dari pola kata antara kata kerja dan kata benda terjadi dalam kata benda verbal "masdar". Kata benda verbal dari bentuk kelima dan keenam sering menimbulkan kebingungan. Misalnya, kata “تدخّل” (bentuk kelima) bisa berupa kata kerja (ikut campur) dan kata benda (gangguan) juga kata "تعاون" (bentuk keenam) bisa berupa kata kerja (untuk membantu) dan kata benda (kerja sama) [18]

Pola (أفعل) bahkan lebih rumit. pola ini menawarkan setidaknya tiga kemungkinan: kata benda, kata sifat atau kata kerja. Kata (أبيض) misalnya berarti putih sebagai putih (anggota ras kulit putih) dan juga dapat memiliki fungsi kata kerja dalam kalimat (ما أبيض وجهه) yang berarti "apa wajahnya putih!" [18]

Pendekatan yang diusulkan berkaitan dengan kata benda, kata kerja dan partikel karena mereka adalah tiga bagian utama dari pidato Arab dan tidak ada kata-kata Arab yang diklasifikasikan di luar bagian-bagian ini dan semua kelas lainnya adalah cabang dari bagian-bagian ini. Dalam studi ini, kami telah berkonsentrasi pada cara-cara untuk benar-benar membedakan antara tiga bagian utama ini yang akan memungkinkan sistem untuk menandai lebih banyak kelas dengan tingkat yang lebih tinggi dalam pekerjaan masa depan.

3. Pekerjaan Sebelumnya

Banyak metode telah diusulkan untuk penandaan kelas kata. Sebagian besar karya menggunakan imbuhan kata-kata dan polanya untuk tujuan ini. Abuleil et al. [1, 2, 3, 4] mengusulkan empat pendekatan untuk pemrosesan bahasa Arab. Pada tahun 1998 [1], mereka membangun leksikon Arab otomatis untuk menandai teks surat kabar Arab. Pada tahun 2002 [2], mereka mengusulkan sistem berbasis aturan yang menggunakan analisis suffix dan analisis pola untuk menganalisis kata benda Arab untuk menghasilkan informasi morfologis mereka berkenaan dengan jenis kelamin dan jumlah. Pada tahun 2004 [3], ia membangun database dan grafik untuk mewakili kata-kata yang mungkin membentuk nama dan hubungan di antara mereka. Pada tahun 2006 [4], mereka menggambarkan sistem pembelajaran yang menganalisis kata benda Arab untuk menghasilkan informasi morfologis mereka berkenaan dengan gender dan angka berdasarkan analisis suffix dan analisis pola.

Banyak teknik berbasis aturan yang diusulkan. Diab et al. [10] merancang sistem penandaan otomatis untuk menandai tag bagian-ucapan dalam teks Arab. Habash et al. [11] mengusulkan penganalisis morfologis untuk tokenizing dan penandaan kata-kata Arab secara morfologis. Khoja [14] mengembangkan sistem penandaan dengan menyisir teknik statistik dengan teknik berbasis aturan. Set tag yang digunakan diekstraksi dari set tag BNC Inggris tetapi dimodifikasi dengan beberapa konsep dari tata bahasa Arab tradisional. Set tag berisi 131 tag yang ditetapkan untuk kata-kata. Sebuah korpus 50.000 kata dari surat kabar Saudi Al-Jazira digunakan untuk melatih sistem penandaan.

Freeman [13] menggambarkan sistem penandaan bagian-bahasa Arab berdasarkan sistem penandaan Brill yang merupakan sistem pembelajaran mesin yang dapat dilatih dengan corpus yang sebelumnya ditandai. Freeman menggunakan satu set tag yang terdiri dari 146 tag yang diekstraksi dari Brown corpus untuk bahasa Inggris. Juga, Lee et al. [23] menggunakan kumpulan kata-kata tersegmentasi secara manual yang tampaknya merupakan bagian dari rilis pertama ATB (110.000 kata). Mereka memperoleh daftar awalan dan sufiks dari korpus ini yang tampaknya ditambah dengan daftar turunan afiks lain yang diturunkan secara manual. Maamouri et al. [15] mempresentasikan sistem penandaan bagian-bahasa untuk bahasa Arab. Para penulis mendasarkan pekerjaan mereka pada hasil analisis morfologis Tim Buckwalter. Sistem penandaan ini diuji pada corpus yang terdiri dari 734 file yang diekstrak dari "Agence France Press" yang dikembangkan oleh Maeda dan Hubert Jin.

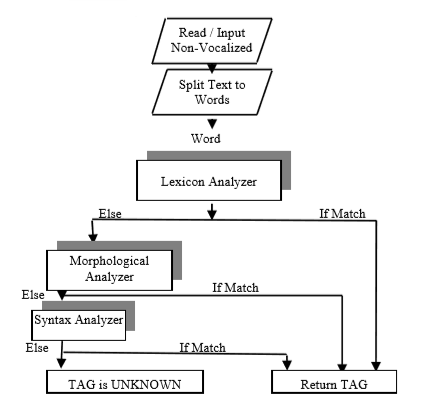
Banyak peneliti mencari metode baru untuk menyelesaikan ambiguitas dalam teks Arab. Marsi et al. [16] mengeksplorasi penerapan pembelajaran berbasis memori untuk analisis morfologis dan penandaan sebagian bahasa Arab tertulis berdasarkan data dari Treebank Arab. Al Shamsi et al. [5] menyelesaikan ambiguitas penandaan teks bagian-ucapan Bahasa Arab melalui penggunaan model bahasa statistik yang dikembangkan dari corpus Arab sebagai Hidden Markov Model (HMM).

Sebagian besar penelitian bahasa Arab diproses dan dianalisis teks non-disuarakan. Tetapi banyak peneliti lain yang memproses dan menganalisis teks yang disuarakan dan membuat aturan tentang vokal pendek (Fatha, Damma, Kasra, Sukun, Tanween-Fateh, Tanween-Damm, TanweenKasir) untuk mengklasifikasikan kata tersebut dan mengidentifikasi kelompok yang menjadi miliknya. Alqrainy et al. [7] mempresentasikan sistem penandaan kata-kata berbasis aturan yang secara otomatis menandai teks Arab yang sebagian disuarakan. Tujuannya adalah untuk menghapus ambiguitas dan mengaktifkan sistem penandaan otomatis cepat yang akurat. Set tag telah dirancang untuk mendukung sistem ini. Desain tag set adalah pada tahap awal penelitian yang terkait dengan anotasi morf-sintaksis otomatis dalam bahasa Arab. Talmon et al. [22] mempresentasikan sistem komputasi untuk penandaan morfologis Alquran untuk tujuan penelitian dan pengajaran. Inti dari sistem ini adalah seperangkat aturan berbasis negara yang menggambarkan fenomena morfologis dan morfologis sintaksis dari bahasa Quran. Sistem saat ini sedang digunakan untuk tujuan pengajaran dan penelitian. Safadi et al. [21] mempresentasikan metode untuk memasok teks Arab yang disuarakan dengan menggunakan metode mesin tanpa pengawasan.

Karya terbaru oleh Chiraz et al. [9] membahas masalah penandaan teks berbahasa Arab dengan tanda vokal. Sistem ini terdiri dari lima agen yang bekerja secara paralel untuk menentukan tag yang cocok untuk setiap kata dalam sebuah kalimat.

4. Metode

Dalam penelitian ini, kami menyajikan sistem penandaan yang mengklasifikasikan kata-kata dalam teks Arab yang tidak disuarakan ke tag mereka. Sistem memproses teks yang tidak disuarakan yang merupakan teks tanpa vokal pendek yang biasanya dihilangkan dari teks Arab seperti koran, buku, dan media. Gambar 1 menunjukkan arsitektur sistem yang diusulkan. Ini terdiri dari tiga tingkat: tingkat pertama adalah penganalisa leksikon yang berisi semua partikel Arab termasuk preposisi, kata keterangan, konjungsi, Partikel interogatif, pengecualian, dan interjeksi. Tingkat kedua adalah penganalisa morfologis yang menggunakan informasi morfologis seperti pola kata dan afiksnya untuk menganggap kelas kata-kata tersebut. Level terakhir adalah penganalisis sintaksis yang terdiri dari dua tahap; tahap pertama tergantung pada kata kunci spesifik yang memberi tahu kami kata yang berurutan, dan tahap kedua adalah teknik penguraian terbalik.



Gambar 1. Arsitektur sistem penandaan

Sistem yang diusulkan membaca teks Arab yang tidak disuarakan dan membaginya menjadi kata-kata yang terpisah, lalu kami mengambil setiap kata dan memasukkannya ke tingkat pertama (Lexicon Analyzer), jika ada kami mengembalikan TAG yang sesuai, jika tidak; kami mentransfer kata ke level kedua (Morphological Analyzer). Setelah memproses kata; jika cocok kami mengembalikan TAG yang dianggap, jika tidak; kami mentransfernya ke tingkat akhir (Sintaks Analisis). Setelah menguji posisi kata, jika TAG ditemukan maka TAG dikembalikan, jika tidak, tidak ada anggapan tentang TAG yang sesuai atau TAG TIDAK DIKETAHUI.

4.1. Lexicon Analyzer

Lexicons adalah jantung dari setiap sistem pemrosesan bahasa alami. Level penandaan awal adalah penganalisa leksikon. Sistem ini memiliki leksikon yang menyimpan semua kata dan partikel dalam bahasa Arab (kata depan, kata keterangan, kata sambung, partikel interogatif, pengecualian, pertanyaan dan kata seru, lihat Lampiran A) (prepositions, adverbs, conjunctions, interrogative particles, exceptions, questions and interjections, see Appendix A). Setiap kata dalam teks bacaan non-disuarakan dieksplorasi dalam leksikon, jika ditemukan; maka TAG yang sesuai dikembalikan. Tetapi jika tidak ditemukan, kami mentransfer kata ke tingkat kedua dari sistem yaitu penganalisa morfologi. Proses analisa leksikon dapat diringkas dengan algoritma berikut:

Begin

read text

tokenization

take word

search for the word in the lexicon

if found then

return the corresponding tag

else

transfer word to the morphological

analyzer

End

Mulai

Baca text

Tokenization

Ambil kata

Mencari kata dalam leksikon

Jika ditemukan maka

Kembalikan tag yang sesuai

Lain

Mentransfer kata ke morfologis

Penganalisa

Akhir

4.2. The Morphological Analyzer

Suatu sistem morfologis adalah tulang punggung dari sistem pemrosesan bahasa alami. Membangun penganalisa morfologi untuk bahasa Arab memiliki motivasi dan tantangan tersendiri yang berbeda dengan yang dibagi untuk semua penganalisa morfologis. Bahasa Arab adalah bahasa yang sangat infleksional dan sangat turunan. Ini masing-masing disebabkan oleh sejumlah besar kemungkinan afiks (terutama awalan dan sufiks), memiliki bahasa Arab, dan sejumlah besar bentuk turunan (pola) dari kata tertentu yang dapat memiliki sistem root yang unik. Karena ada banyak dan beragam aturan morfologi Arab, banyak penelitian dapat menerapkan lebih dari satu pendekatan menggunakan aturan ini. Untuk banyak operasi morfologi Arab klasik, ahli bahasa memiliki cara berbeda untuk bekerja dengan operasi yang sama.

Ada beberapa tanda dalam bahasa Arab yang menunjukkan apakah kata itu kata benda atau kata kerja. Salah satu tanda adalah pola kata, beberapa pola digunakan dengan kata kerja dan lainnya digunakan dengan kata benda. Tetapi ketika kita berurusan dengan kata-kata yang tidak disuarakan, banyak kata yang ambigu karena polanya digunakan dengan kata kerja dan kata benda. Bagian-of-speech dari sebuah kata juga dapat ditemukan dengan menggunakan imbuhan. Beberapa afiks digunakan dengan kata kerja saja, beberapa digunakan dengan kata benda saja dan beberapa digunakan dengan kata kerja dan kata benda. Banyak peneliti mendaftar dan mendefinisikan beberapa awalan dan sufiks yang mengidentifikasi kelas kata yang diberikan untuk disuarakan atau tidak. Karena sebagian besar kata bahasa Arab adalah trilateral, para ahli morfologi menganggap bahwa asal-usul kata-kata Arab adalah tiga huruf. Untuk alasan ini, jika kita ingin menyeimbangkan kata untuk mengetahui asal kata itu (imbuhan: awalan, infiks, dan akhiran) kita harus menghadapi huruf asli kata tersebut dengan huruf kata (فعل) jika panjang kata lebih besar dari tiga huruf kita menghadapi huruf asli untuk kata seimbang dengan huruf kata (فعل) dan huruf-huruf tambahan dihadapkan dengan pengucapan seperti (انتظار ... افتعال) ,(استشعار ... استفعال) ,(المنازل ... المفاعل) seperti pada contoh berikut:

**انتظار هم ا ن ت ظ ا ر ه م**

**افتعالهم ا ف ت ع ا ل ه م**

Afiks selalu merupakan bagian dari kata "سألتمونيها" yang datang dalam kata di empat posisi [25]: sebelum 'fa' kata (قبل فاء الكلمة) yang disebut awalan, setelah 'fa' dari kata (بعد فاء الكلمة) yang disebut infix, setelah 'ayn' dari kata (بعد عين الكلمة) yang disebut infix, dan setelah 'lam' dari kata (بعد لام الكلمة) yang disebut suffix. Prefiks dan sufiks dapat digunakan tidak hanya untuk mengekstrak informasi seperti fitur tense dan subjek / objek, tetapi juga untuk membantu mengidentifikasi jenis token (kata benda atau kata kerja). Ini karena beberapa awalan / sufiks melekat pada jenis token tertentu seperti yang ditunjukkan pada Gambar 2 dan karenanya dapat digunakan dalam proses disambiguasi. Terkadang satu afiks dapat menentukan tag kata: misalnya, jika awalannya adalah ( ال – التعريف maka kata itu adalah kata benda [17] ).

Tubuh kata adalah bagian utamanya. Ini disebut batang kata (ساق/جذع الكلمة). Itu adalah bagian dalam yang dikelilingi oleh beberapa awalan dan sufiks. Jadi sistem yang diusulkan berkonsentrasi untuk menemukan batang kata sebagai langkah pertama. Kemudian, dari batang kami mengasumsikan awalan, sufiks dan infiks yang dapat memetakan kata ke TAG yang telah ditentukan sebelumnya. The Morphological Analyzer menggunakan pendekatan sederhana untuk membagi kata Arab menjadi tiga bagian:

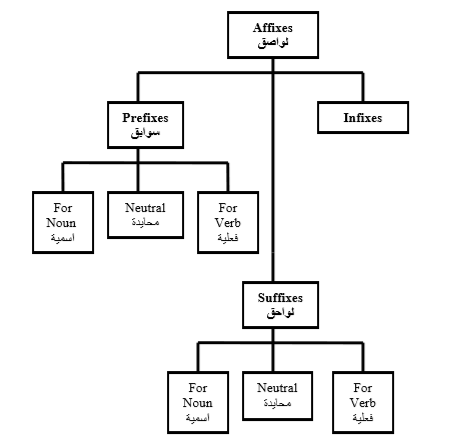
*Prefix:* Terdiri dari sebanyak lima awalan bersambung atau bisa nol (Lampiran B).

*Stem:* Ini terdiri dari akar dan pola morfem yang berasal dari kata benda dan pola kata kerja [25] (Lampiran D).

*Suffix:* Terdiri dari sebanyak tiga sufiks gabungan atau bisa nol (Lampiran C). Misalnya, kata "syktbwnha سيكتبونها" akan dianalisis sebagai berikut:

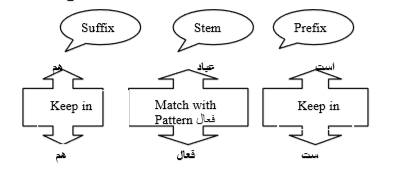
Prefix Stem Suffix

Sy سي ktb كتب wnhA ونها



Gambar 2. Afiks kelas khusus

Kami telah mengekstrak awalan umum terpanjang dari kata yang diberikan dengan membandingkan kata dengan awalan pada Tabel 4 (lihat lampiran B). Sufiks umum terpanjang diekstraksi dari kata yang diberikan dengan membandingkan kata dengan sufiks pada Tabel 5 (lihat lampiran C), kemudian kami membandingkan huruf yang tersisa dengan pola batang yang ada (tercantum dalam lampiran D) dan mengabaikan tiga huruf kata yang sesuai فعل sebagai root dan kemudian ambil infiksnya. untuk exmple, untuk kata استعبادهم (gambar 3) pertama-tama kita mengekstrak awalan umum terpanjang (است) setelah dicocokkan dengan awalan pada tabel 4, kita mengekstrak sufiks (هم) setelah dicocokkan dengan sufiks pada tabel 5. kemudian kita membandingkan sisanya. huruf dengan yang ada dengan patters batang (lihat lampiran D), itu akan dicocokkan dengan pola batang (فعال), maka kita mengabaikan huruf kata فعل yang merupakan root dan mengekstrak huruf yang tersisa sebagai infixes.



Gambar 3. Contoh.

Kami percaya bahwa informasi yang diberikan oleh afiks gramatikal akan berguna tetapi tidak cukup untuk menentukan klasifikasi kata yang tepat dalam dua kategori utama kata benda dan kata kerja. Awalan, sufiks atau infiks tertentu datang dengan kelas kata tertentu, jadi kami mencoba untuk mengumpulkan aturan awalan, sufiks, dan infiks tata bahasa kami sendiri yang Mengidentifikasi kata kerja dan kata benda.

Penganalisa morfologis yang diusulkan membangun daftar awalan yang berbeda, daftar akhiran yang berbeda, daftar infiks yang berbeda, dan daftar hubungan antara awalan, sufiks dan infiks. Semua daftar ini memungkinkan sistem mengidentifikasi kelas kata yang diberikan.

Kami melakukan studi morfologi intensif dan berkonsentrasi pada afiks yang membangun pola unik kata benda dan kata kerja untuk mengenali apa yang membuat afiks. Sebagai hasilnya, kami mengembangkan daftar aturan yang mengenali prefixes (awalan), sufiks, infiks, dan hubungan di antara mereka untuk mengidentifikasi kelas kata. Aturan-aturan ini adalah:

Aturan 1 : Awalan berikut (atau bagian dari awalan) memetakan kata ke kelas NOUN: AL, FL, LL, M,Afiks preposisi (b, k) (ال التعريف, فل, لل, م,زواتدأحرف الجرب,ك)

Aturan 2 : Sufiks berikut (atau bagian dari sufiks) memetakan kata ke kelas NOUN: T, AT, AA (ة,ات,اء)

Aturan 3 :Sufiks berikut (atau bagian dari sufiks) memetakan kata ke kelas KATA NOUN dengan kondisi tidak ada huruf yang tidak sempurna (المضارعة أحرف) : WN, YN, AN, Y: (ون, ين, ان,ي)

Aturan 4 : Infiks berikut memetakan kata ke kelas KATA BENDA dengan kondisi memuaskan posisi yang sesuai dalam pola batang (stem) ditentukan antara tanda kurung: A, Y, AW, AWY (setelah ayn dari kata) (بعد عين الكلمة) ا, ي, و, او, اوي WA (setelah fa 'dari kata tersebut) (بعد فاء الكلمة) وا

Aturan 5 : prefixes berikut (atau bagian dari awalan) memetakan kata ke kelas VERB: Y, N, A ’, Future S (ي, ن, أ. س الاستفبال)

Aturan 6 : Awalan berikut (atau bagian dari awalan) memetakan kata ke kelas VERB dengan ketentuan aturan yang memetakan kata ke kata benda : A (ا)

Aturan 7 : Sufiks berikut memetakan kata ke kelas VERB: Pembukaan T (ت المفتوحة)

Kita dapat melihat dari aturan-aturan ini bahwa sistem harus mengekstrak awalan, akhiran, batang dan pola batang dari kata untuk menggunakan aturan-aturan ini. Setelah aturan dipenuhi; kelas kata diidentifikasi. Misalnya, dalam kata ALMADRASA (المدرسة); kami memiliki bagian dari awalan adalah AL (ال التعريف) dan bagian lainnya adalah M (م) dan keduanya memetakan kata ke kelas kata benda dengan aturan 1. suffix adalah T (ة) yang memetakan kata ke kelas kata benda oleh aturan 2. Setelah mengekstraksi batang dan merencanakan polanya, kita dapat mengenali posisi infiks setelah fa 'atau' ayn 'dari kata tersebut (بعدفاع أو عين الكلمة). sebagai contoh, dalam kata ALMADROSA, kita memiliki infiks W (و) setelah 'ayn' dari kata (بعدعبن الكلمة) yang memetakan kata ke kelas kata benda dengan aturan 4. Aturan sebelumnya memberi kita peluang tinggi untuk menganggap kelas kata yang diberikan, tetapi beberapa kata mungkin tidak memiliki imbuhan untuk membimbing kita ke TAG yang sesuai, dalam kasus seperti itu kita meneruskan kata tersebut ke tahap akhir yaitu penganalisis sintaks. Alat analisis morfologi dapat diringkas dengan algoritma berikut:

Begin

take word;

extract the stem;

extract all affixes with defined positions

test the seven rules

if one of them satisfied then

return the corresponding tag

else transfer word to the syntax analyzer

End

Mulai

Ambil Kata

Ektraks Stem

Ekstrak semua afiks dengan posisi yang ditentukan

Uji ketujuh aturan tersebut

Jika salah satu dari mereka puas maka

Kembalikan tag yang sesuai

lain mentransfer kata ke analisa sintaksis

Akhir

4.3. Penganalisis Sintaks

Fase ini digunakan jika kata itu tidak memiliki afiks untuk memandu penganalisa morfologis. Analisis sintaksis mungkin merupakan aspek yang paling banyak dipelajari dan paling dipahami dalam pemrosesan bahasa. Kami menggunakan dua aturan untuk memetakan kata yang mendua ke TAG yang sesuai: konteks kalimat dan parsing terbalik.

4.3.1. Konteks Kalimat

Tahap ini tergantung pada hubungan dengan kata-kata yang berdekatan dan terkait dalam frasa atau kalimat. Dalam bahasa Arab, posisi kata dalam kalimat adalah indikator yang baik untuk mengidentifikasi kata benda dari kata kerja.

Preposisi (حروف الجر) dan interjections (kata seru) (حروف النداء) selalu diikuti oleh kata benda seperti dalam kata: “fe almadrasa” (فى المدرسة) dan dalam kata “ya Mohamed” (يا محمد). Beberapa kata selalu diikuti oleh kata benda seperti kata-kata: السيد, الشيخ, المملكة, الر ئيس [6].

4.3.2. Reverse Parsing

Penganalisis morfologis berhasil memecahkan kata-kata yang hampir tidak bersuara dalam corpus Arab, tetapi ada beberapa kata yang memiliki struktur ambiguitas yang mencegah penganalisa morfologis menebak kelasnya. Misalnya kata KTB (كتب) bisa berupa kata kerja yang berarti "menulis" atau kata benda yang berarti "buku".

Dalam penelitian ini kami telah mengembangkan tata bahasa bebas konteks bahasa Arab untuk menentukan kelas kata-kata jenis ini. Misalnya, dalam kalimat:

(ذهب الولد الى المدرسة)

(؟؟؟ اسم حرف جر اسم)

(thahaba alwalado ela almadrasate)

(??? KATA BENDA preposisi KATA BENDA)

Aturan dengan (kata kerja, kata benda, preposisi, dan kata benda) dihitung dan kelas kata kerja (thahaba ذهب) digandakan sebagai kata kerja. dalam kalimat "ذهب الولد الى المدرسة", kata ذهب bersifat ambigu dan telah gagal untuk mengidentifikasi TAG-nya. Sementara kata-kata lain dalam kalimat telah berhasil mengidentifikasi TAG-nya. Kata الى adalah KATA BENDA yang memenuhi aturan1, kata المدرسة adalah partikel preposisi yang disimpan dalam leksikon, dan kata itu KATA BENDA yang memenuhi Aturan. Ketika kami membandingkan kalimat di atas dengan aturan bahasa Arab yang tersimpan, kami menemukan bahwa aturan (kata kerja فعل, kata benda اسم, prepostion حرف جر, kata benda اسم) cocok dengan kalimat. Ketika kami mengabaikan kata ذهب dan menelusuri aturan bahasa Arab dan kalimat yang cocok sebagai alternatif; kita bisa menebak dan mengembalikan TAG yang cocok dengan kata ambigu yang sesuai, jadi kami mengembalikan kata kerja (فعل) TAG ke kata ambigu ذهب. Jumlah aturan yang digunakan dalam parsing terbalik adalah 10 aturan bahasa Arab yang paling sering digunakan yang berisi VP atau NP atau keduanya. Aturan-aturan ini mencakup urutan sebagai berikut: kata benda kata benda, kata benda kata benda, kata benda benda benda, nomina benda benda nomina, kata benda benda benda nomina, nomina . Jika ada lebih dari satu aturan yang sesuai dengan kalimat untuk dianalisis, kita mengabaikan kata yang merupakan kata yang tidak dianalisis.  
Kami dapat merangkum proses parsing terbalik dengan algoritma berikut:

Begin

List sequence of tags corresponds to each word

Ignore the tag of ambiguity word

Compare a sequence of tags with a stored cfg

When one grammar rule matched

Trace the sentence with the matched rule

Return the tag of the ambiguous word

End

Mulai

Daftar urutan tag sesuai dengan setiap kata

Abaikan tag kata ambiguitas

Bandingkan urutan tag dengan cfg yang disimpan

Ketika satu aturan tata bahasa cocok

Lacak kalimat dengan aturan yang cocok

Kembalikan tag kata yang mendua

Selesai

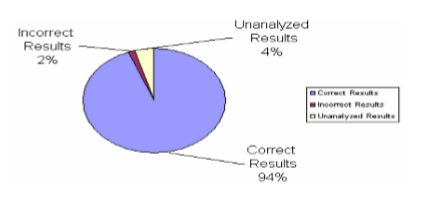
5. Eksperimen dan Evaluasi

Kami telah menguji keakuratan pendekatan yang diusulkan dengan menggunakan kumpulan data yang terdiri dari 2355 kata-kata Arab yang tidak disuarakan dalam 10 artikel surat kabar yang dipilih secara acak. Tabel 1 menunjukkan jumlah kata dalam setiap artikel, jumlah TAG tebakan yang berhasil dipetakan ke kata-kata dan persentase TAG tebak yang berhasil. Dapat dilihat dari Tabel 1 bahwa sistem berhasil menganalisis 2211 kata dan memetakannya ke TAG yang sesuai, dan gagal menganalisis 144 kata, yaitu, ia memperoleh tingkat keberhasilan 94%. Beberapa kata yang tidak dianalisis memberikan hasil yang salah dan yang lainnya tidak dianalisis; persentase hasil yang salah dan kata-kata yang tidak dianalisis dapat diukur dengan mengingat dan presisi. Kami telah menghitung ketepatan dan penarikan kembali dari pendekatan yang diusulkan menggunakan rumus berikut:

Tabel 1. Akurasi sistem klasifikasi kata

|  |  |  |  |
| --- | --- | --- | --- |
| Articles | Number of Words | Number of Successful TAGs | Percentage of Successful |
| Article 1 | 240 | 226 | 94% |
| Article 2 | 173 | 164 | 95% |
| Article 3 | 254 | 238 | 94% |
| Article 4 | 396 | 369 | 93% |
| Article 5 | 127 | 119 | 94% |
| Article 6 | 147 | 138 | 94% |
| Article 7 | 208 | 198 | 95% |
| Article 8 | 361 | 339 | 94% |
| Article 9 | 282 | 263 | 93% |
| Article 10 | 167 | 157 | 94% |

Hasil dari tahap ini diberikan pada Tabel 2. Dapat dilihat dari Tabel ini bahwa sistem memperoleh sekitar 98% keseluruhan presisi untuk kata-kata yang dianalisis. Tabel ini juga menunjukkan bahwa sistem tersebut memperoleh sekitar 96% penarikan keseluruhan untuk kata-kata yang dianalisis dan tidak dianalisis. Kata-kata yang tidak cocok dengan TAG yang benar diabaikan. Sistem klasifikasi kata yang diusulkan salah klasifikasi 4% dari kata-kata yang diuji, memberikan 2% hasil yang salah dari kata-kata yang dianalisis dan berhasil menganalisis dan mendapatkan hasil yang benar 94% dari kata-kata yang diuji, seperti yang ditunjukkan pada Gambar 4



Gambar 4. Hasil akhir.

6. Kesimpulan

Kami telah merancang dan menerapkan sistem klasifikasi berbasis aturan untuk menyelesaikan masalah anotasi teks Arab yang tidak disuarakan dengan tag. Kami menyimpan semua partikel dan kata-kata tetap dalam leksikon, dan kami telah mengungkapkan cara menggunakan penganalisa morfologis untuk tokenisasi dengan mengekstraksi awalan dan sufiks serta mengekstraksi infiks dari pola batang, kemudian menelusuri aturan sampai salah satu dari mereka cocok.

Kami telah menunjukkan bagaimana menggunakan konteks kalimat dan struktur bahasa Arab untuk membangun parsing terbalik untuk memecahkan sebagian besar kata yang tidak dianalisis dalam penganalisa morfologis. Ketiga analisis dalam sistem yang diusulkan dapat digunakan dengan sukses untuk menentukan persentase kelas kata yang tinggi.

Sebagian besar karya sebelumnya menggunakan analisis awalan dan sufiks, tetapi pendekatan yang diusulkan memiliki keuntungan menggunakan analisis awalan, sufiks, dan infiks. Menambahkan analisis infiks membantu dalam memecahkan banyak kasus ambigu dari kata benda dan kata kerja yang memiliki awalan dan sufiks yang serupa.

Posisi kata dalam kalimat adalah indikator yang baik dalam mengidentifikasi kata benda. Banyak peneliti menggunakan fenomena ini untuk membangun aturan untuk membantu mengidentifikasi kata benda dalam teks seperti pada [22] dan yang lain menggunakannya untuk mengidentifikasi nama pribadi dalam teks seperti sistem QARAB [12]. Sebaliknya, pendekatan kami mampu memberikan cakupan penuh untuk mengidentifikasi kata benda dan nama pribadi. Pendekatan kami juga menggunakan fenomena ini untuk membangun teknik baru yaitu teknik penguraian terbalik yang memindai tata bahasa Arab yang tersedia untuk mendapatkan kelas kata ambiguitas tunggal dalam kalimat berdasarkan pada posisinya.

Tabel 2. Akurasi sistem klasifikasi kata

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Article | No. of Correct Result | No. of Incorrect results | No. of Misclassified Words | Precision | Recall |
| Article 1 | 226 | 4 | 10 | 98% | 96% |
| Article 2 | 164 | 4 | 5 | 98% | 97% |
| Article 3 | 238 | 4 | 12 | 98% | 95% |
| Article 4 | 369 | 8 | 19 | 98% | 95% |
| Article 5 | 119 | 4 | 4 | 98% | 97% |
| Article 6 | 138 | 3 | 6 | 98% | 96% |
| Article 7 | 198 | 4 | 6 | 98% | 97% |
| Article 8 | 339 | 8 | 14 | 98% | 96% |
| Article 9 | 263 | 6 | 13 | 98% | 95% |
| Article 10 | 157 | 3 | 7 | 98% | 96% |

7. Pekerjaan Masa Depan

Banyak teknik telah diusulkan untuk menandai Bahasa Inggris dan korpora bahasa Eropa lainnya. Salah satu teknik yang dikembangkan adalah teknik berbasis aturan dan semua teknik lain diperluas untuk itu. Teknik berbasis aturan adalah teknik yang kami gunakan dalam sistem kami, sehingga kami dapat memanfaatkan dari aturan kami di penganalisa morfologi untuk membangun teknik baru seperti model statistik atau analisis semantik untuk memetakan kata yang diberikan ke TAG yang sesuai.

Referensi

[1] Abuleil S. and Evens M., “Discovering Lexical Information by Tagging Arabic Newspaper Text,” CSAM, Illinois Institute of Technology, Chicago, 1998.

[2] Abuleil S., Alsamara K., and Evens M., “Acquisition System For Arabic Noun Morphology,” in Proceedings of the Workshop on Computational Approaches to Semitic Languages, USA, pp. 1-8, 2002.

[3] Abuleil S., “Extracting Names from Arabic Text for Question Answering Systems,” MIS Department, Chicago State University, 2004.

[4] Abuleil S. and Alsamara K., “Enhance the Process of Tagging and Classifying Proper Names in Arabic Text,” in Proceedings of the International Arab Conference on Information Technology (ACIT'2006), Jordan, pp. 43, 2006.

[5] Al Shamsi F. and Guessoum A., “A Hidden Markov Model-Based POS Tagger for Arabic,” in Proceedings of the 8th International Conference on the Statistical, 2006.

[6] Al-Shalabi R. and Kanaan G., Constructing an Automatic Lexicon for Arabic Language, Yarmouk University, Jordan, 2004.

[7] Alqrainy S. and Ayesh A., “Word-Class Tagger and Tagset Design for Vocalized Arabic Text,” in Proceedings of the 2nd Jordanian International Conference on Computer Science and Engineering (JICCSE 2006), Jordan, pp. 278-283, 2006.

[8] A Web of Morphology, http://angli02. kgw.tuberlin.de/call/webofdic/morph.html, 2007.

[9] Chiraz Z., Aroua T., and Mohamed A., “A Multi Agent System for POS-Tagging Vocalized Arabic Texts,” International Arab Journal of Information Technology (IAJIT), vol. 4, no. 4, pp. 322-329, 2007.

[10] Diab M., Hacioglu K., and Jurafsky D., “Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks,” Linguistics Department, Stanford University, 2004.

[11] Habash N. and Rambow O., “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, 2005.

[12] Hammo B., Abu-Salem H., and Lytinen S., “QARAB: A Question Answering System to Support the Arabic Language,” in Proceedings of Workshop on Computational Approaches to Semitic Language (ACL), Philadelphia, pp. 55-65 2002.

[13] Freeman A., “Brill's POS Tagger and a Morphology Parser for Arabic,” Department of Near Eastern Studies, Michigan, USA, 2001.

[14] Khoja S., “APT: Arabic Part-of-Speech Tagger,” Computing Department, Lancaster University, Lancaster, 2003.

[15] Maamouri M. and Cieri C., “Resources for Arabic Natural Language Processing at the Linguistic Data Consortium,” in Proceedings of the International Symposium on Processing of Arabic language Faculté des Lettres, Tunisia, 2002.

[16] Marsi E. and Soudi A., Memory-based Morphological Analysis Generation and Part-ofSpeech Tagging of Arabic, Tilburg University, 2006.

[17] Mohamed A., “A Large-Scale Computational Processor of the Arabic Morphology and Applications,” Master Thesis, Faculty of Engineering, Cairo University, Egypt, 2000.

[18] Mol V, “The Semi-Automatic Tagging of Arabic Corpora,” Katholieke University, ILT, 2004.

[19] Mustafa S. and Awwad S., “Arabic Word Class Tagging Based on the Analysis of Affix Structure,” in Proceedings of the International Arab Conference on Information Technology (ACIT'2006), Jordan, pp. 145-145, 2006.

[20] Nachum D., “Part of Speech Tagging,” Seminar in Natural Language Processing and Computational Linguistics, USA, 2007.

[21] Safadi H., Dakkak O., and Ghneim N., Computational Methods to Vocalize Arabic Texts, Syria, 2006.

[22] Talmon R. and Wintner S., Morphological Tagging of the Qur’an, University of Haifa, Israel, 2001.

[23] Young-Suk L., Papineni K., and Roukos S., “Language Model Based Arabic Word Segmentation,” in Proceedings of the Annual Meeting on Association for Computational Linguistics, Japan, pp. 399- 406, 2003.

[24] اB\آ6Pر M\ اBQ¡5¡ M6b5 ، "اBt5\UL اB5@ M|5 اB5¢Wف "، دار ١٩٧١ اBse{ اBQWcb{، .

[25] اB55\آ6Pر ز55 آNS55L اB¦P55}@، " اB¡وا55\ O55 اB55¢b¨ O55 اB|§55 { ١٩٧١ اBQWcb{"، دار اBtQWO{ اBNSQb{، .

[26] اc اBA©Nع اB¢A|، "أcsb{ اªv5tNء و اªOQ5Nل و اBt5 ¢Nدر"، دار ١٩٩٩اB}64 و اBP«N  اBAPSb{، .

[27] ONرس St\ Mb@ دار ،" M| اB¢Wف S se¬ O5 اB56Q| اB5ا] " ، .٢٠٠٠ اBR}W B|©NM{ و اBsuW و اB6Pز­، MtNن، اªردن،

Appendix A

Table 3. Lexicons

Table 3. Lexicons.

S، اB@، إB@، 6@، ّ6@، M|@، M، O، رّب، S، N²N Ss، U±، M\ا،

Wوف اBW Prepositions Wوف اBQ© Lc و، «، «ّ، أو، أم، Conjunctions ت Wوف اBsR ، ،NS ،B ،NtB ،B Negation Particles yQ، c|@، إي، أfL، آّ|N Wوف اBPاب Answering Particles WbR] أي Wف Explanation Particle إن، إذSN، BP، BP، BPSN، أّSN، BّtN Wوف اBuWط Conditional Particles \s] و ab] هّ|N، أّBN Wوف أ، أSN Wوف اBQWض أ ّن ، أن، آ Wوف S¢\ر { Verbal Noun Particles vPف Wف اv6ANل Future Particle إّن، \ Wوف اB6Pآb\ Emphasis Particles هL Wف اv6ReNم Interrogative Particles st] Wف bB Wishing Particle و إ²RNق fّW] Wف LّQB ،LQB N، أN Wوف اBs\اء Interjections و yN LQRBNc duS إّyN إن Wف Wbtµ W66S إّBN Wف اv6¶sNء Exceptive Particle إّSN إن اBuW·b{ و SN زا\ة آzّytN آzن اBuW·b{ و SN زا\ة آtN اB}Nف Wف fW و SN زا\ة رّctN ر ّب Wف fW و SN زا\ة اBt6}| WNtµ أyN، y First Person Pronouns أy، أy6tN، أy6، أy6 µtNW اBt¦N·4 Second Person Pronouns هP، ه، ه، هtN µtNW اB§N4 Third Person Pronouns ها، ذB¸، هي، ]|¸، هان، هN]Nن، هºء، أوB¹¸، هsN، هsNB¸، ه» أvtNء ا¼²Nرة Demonstrative Pronouns {BP½PS اBي، اB6، اB|ان، اB|6Nن، اB أvtNء Relative Pronouns OPق، ]، أSNم، وراء، tb، ²tNل، U|، إ«W ¾Wوف اBt}Nن Nouns of Place و، vNM{، Pم، ²eW، vs{، MNم، زSNن، ،b أوان، c}Wة، µPة، Bb|{، SNء، Mub{، ¿\وة ¾Wوف اB¡SNن Nouns of Time دون، cb، وvÀ، Ms\، L، cQ\ ¾Wوف زSNن أو S}Nن Nouns of Place or Time أن، إّّن، B}ّ، آzّن، BQL، Bb إ ّن و أUPا]eN Inna and its sisters آNن، أS@، أµ@، أ½Â، Bb، cNت، SNBÁ، ¸SNcWح، SN اyR آNن و أUPا]eN Kaana and its sisters

Appendix B

Table 4. Common prefixes.

No. of Letters Prefixes 5 " اBt " 4 " bB " "، "bv" ،"bO"، "6B"، "6v"، "6O"، "sB" ،"sv"، "sO"، "BNc" ،"اv" ،"y"، "ONل" ،"cNل" ،"]" ،"" ٣ "وال" ،"B|" ،"BÄ" ،"vb"، "آNل" ،"S" ،"B" ،"B" ،"BL"، "ال" ،"v" ،"v" ،"v" ،"vz" ٢ "B"، "" ،"y"، ""، "]" "ا"،" ١ ي "،"ن"،"أ"،"ت "،"ل"،"ف "

Appendix C

Table 5. Common suffixes.

No. of letters Suffixes "ا]etN" 5 "وهtN"، "]etN" 4 "وهN"، "]tN" ،"]e" ،"]sN" ،"]s" ،"وه" ٣ ،"ات" ،"آ" ،""، "ه" ،"آ" ،"]" ،"ه" ،"]" ٢ "هN"، "yN" ،"ون" ،"وك" ،"]¸"، "ان" ،"وا" ا""، "ي" ،"ت" ،"»" ،"ن" ،"ك" ،"ة" ١

Appendix D

Pola batang atau kata-kata dengan infiks setelah menghilangkan awalan dan sufiks yang diekstraksi dan digunakan dalam penganalisa morfologi sistem kami, dapat diabstraksikan dengan: §MLØŒ ONMPÙ „ØŒ OPØ§MbLØŒ OQ | LØŒ OQNBL.

Ahmad Al-Taani adalah profesor kecerdasan buatan di Universitas Yarmouk, Yordania. Ia menerima gelar Sarjana sains di bidang komputer pada tahun 1985 dari Universitas Yarmouk, Yordania. Ia menerima gelar Magister sains di bidang rekayasa perangkat lunak dari Universitas Nasional, Amerika Serikat pada tahun 1988. Ia menerima gelar PhD dalam bidang komputer dari University of Dundee, Inggris pada tahun 1994. Minat penelitiannya meliputi pemrosesan gambar, pemrosesan bahasa Arab, terjemahan mesin, terjemahan bahasa Arab dan web Arab klasifikasi halaman.

Salah Abu Al-Rub adalah pengembang dan pengelola oracle di Bursa Efek Amman. Ia menerima gelar Sarjana ilmu komputer pada tahun 2005 dari Universitas Yarmouk, Yordania. Dia menerima gelar Master of computer science pada tahun 2007 dari Yarmouk University, Jordan. Minat penelitiannya termasuk stemmer Arab dan penandaan paruh-bahasa Arab.