ISSN Print : 2085-1588
ISSN Online : 2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index

Dampak Kelas Kata Bahasa Arab Terhadap Hasil Mesin Penerjemah Berbasis Statistik

Rahmat Izwan Heroza
Sistem Informasi Fakultas Ilmu Komputer Universitas Sriwijaya
e-mail: rahmatheroza@unsri.ac.id

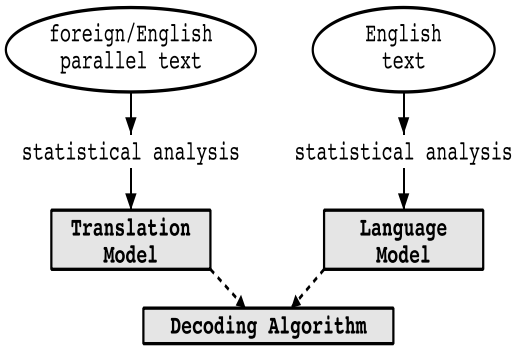
Abstrak

Tulisan ini meneliti dampak kata yang dapat memiliki beberapa kelas kata terhadap hasil mesin penerjemah berbasis statistik dalam menerjemahkan dokumen berbahasa Arab ke dalam bahasa Indonesia. Penelitian ini mengungkap bahwa kalimat yang memiliki kata yang memiliki beberapa makna diterjemahkan tidak sesuai dengan kelas katanya. Penelitian ini akhirnya mengusulkan solusi untuk mengurangi kesalahan pemilihan makna tersebut. Hasilnya adalah tingkat efektifitas sebesar 20% atau sebanyak 5 dari 25 kalimat yang mengandung kata yang dapat memiliki beberapa kelas kata bisa diterjemahkan dengan baik menggunakan mesin penerjemah berbasis statistik yang telah mengimplementasi solusi yang diusulkan dalam penelitian ini..

***Kata kunci** : mesin translasi berbasis statistik bahasa Arab, kelas kata, Moses, PoS tag*

1. PENDAHULUAN

Mesin translasi berbasis statistik adalah paradigma mesin translasi dimana hasil translasi diperoleh dari model statistik yang dibentuk dari proses analisis dua buah dokumen yang sama dalam bahasa yang berbeda [3]. Sebuah mesin translasi berbasis statistik memiliki dua komponen utama yaitu *training pipeline* dan *decoder*. Training pipeline terdiri dari translation model dan language model (Gambar 1) [2].



Gambar 1. Komponen Mesin Translasi Berbasis Statistik

Translation model adalah sebuah model yang memetakan frase sumber kepada frase target beserta peluang translasinya. *Translation model* adalah sumber pengetahuan utama pada *decoder* sebuah mesin translasi. *Decode* menggunakan *translation model* untuk menentukan bagaimana mentranslasikan suatu frase input dari satu bahasa kepada frase output dari bahasa yang lain. Contoh data yang terdapat pada translation table dapat dilihat pada tabel 1

ISSN Print : 2085-1588
ISSN Online : 2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index

Table 1. Contoh *Translation Table*

mn bEd *lk ||| setelah itu ||| 0.333333 0.0267916 0.2 0.00388078 2.718

Language model adalah sebuah model statistik yang dibentuk dari dokumen berbahasa target yang digunakan oleh *decoder* untuk memastikan bahwa hasil translasi yang dihasilkan memiliki susunan kalimat yang baik. Model statistik yang terbentuk merupakan peluang beberapa kata muncul secara berurutan (*N-gram*). Umumnya *language model* yang digunakan dalam sebuah aplikasi hanya menghitung peluang satu (*unigram*), dua (*bigram*) atau tiga kata (*trigram*) yang muncul secara berurutan.

Decoder adalah sebuah komponen yang berfungsi untuk menentukan kalimat dalam bahasa target yang sesuai dengan kalimat sumber yang memiliki nilai tertinggi berdasarkan perhitungan *translation model* dan *language model*. Masing-masing potongan frase dari kalimat masukan akan dicari kemungkinan terjemahannya pada tabel frase sehingga terbentuklah hipotesa-hipotesa hasil translasi. Dengan algoritma tertentu, *decoder* akan memilih hasil translasi yang paling baik dari hipotesa-hipotesa yang ada. Diantara algoritma-algoritma decoding adalah *beam search*, A* *search* dan *Greedy hill-climbing*.

Bahasa Arab adalah bahasa utama yang dipakai oleh negara-negara yang berada di jazirah Arab. Bahasa Arab terdiri dari dua jenis yaitu bahasa Arab klasik (*classical Arabic*) dan bahasa Arab Standar Modern (*Modern Standard Arabic*). Secara keseluruhan, terdapat 28 konsonan dan 3 fonem vokal dalam bahasa Arab. Sebagian besar stem bahasa Arab didasarkan pada akar dari dua atau tiga konsonan antara yang vokal yang dimasukkan. Secara umum, akar konsonan membawa makna semantik dari kata sementara huruf vokal dan susunan vokal-konsonan mencerminkan infleksi kata dan PoS nya.

Penulis tertarik untuk meneliti dampak yang akan terjadi ketika metode statistik digunakan dalam menerjemahkan dokumen berbahasa Arab ke dalam bahasa Indonesia. Hal ini dikarenakan banyak kata dalam bahasa Arab yang memiliki banyak makna. Makna ini tergantung dengan kelas kata tersebut dalam kalimat. Sebagai contoh, kata yang terdiri dari susunan huruf ‘ba’, ‘syin’ dan ‘ro’ selain dapat dibaca “basyaro” yang berarti “manusia”, juga dapat dibaca “basysyir” yang berarti “kabarkanlah”. Perbedaan cara membaca ini disebabkan oleh perbedaan kelas kata tersebut dalam kalimat.

Di akhir tulisan, penelitian ini menguji salah satu usulan solusi yang bisa digunakan untuk mengatasi dampak yang muncul akibat perbedaan kelas kata yang berupa buruknya hasil terjemahan mesin translasi berbasis statistik dari bahasa Arab ke bahasa Indonesia.

Penelitian ini diharapkan dapat memberikan kontribusi pada bidang mesin penerjemah berbasis statistik dari bahasa Arab ke bahasa Indonesia dalam mengetahui faktor-faktor yang menyebabkan buruknya hasil terjemahan yang dihasilkan. Sehingga penelitian berikutnya dapat dilakukan untuk menangani secara khusus faktor-faktor tersebut agar dihasilkan hasil terjemahan yang lebih baik

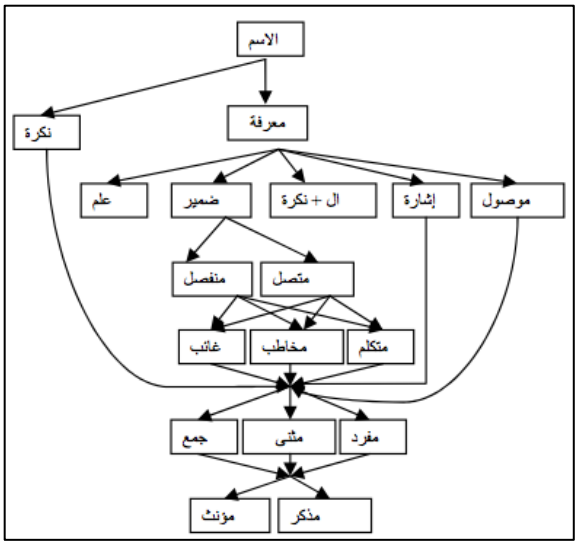
2. KELAS KATA BAHASA ARAB

Salah satu pembagian kelas kata dalam bahasa Arab yang diusulkan adalah [1]:

1. NOUN: berupa kata benda

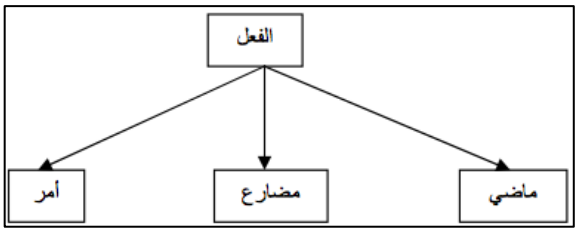
ISSN Print : 2085-1588
ISSN Online : 2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index

2. VERB: berupa kata kerja
3. PAR: berupa partikel

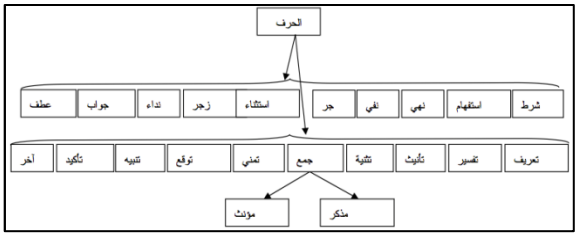


Gambar 2. Pembagian NOUN

NOUN dapat berupa nama atau kata yang mendeskripsikan seseorang atau benda. NOUN dapat *definit*e atau *indefinit*e dan dapat dikategorikan berdasarkan posisi, jumlah, jenis kelamin dan kedudukan kata. Gambar 2 memperlihatkan pembagian NOUN dalam bahasa Arab. VERB adalah kata yang menunjukkan aksi dan dapat digabungkan dengan beberapa partikel. VERB dapat menunjukkan kepada aksi masa lampau, masa sekarang, atau masa depan. Gambar 3 memperlihatkan pembagian VERB dalam bahasa Arab. PAR adalah kata-kata yang tidak termasuk ke dalam NOUN ataupun VERB. Umumnya PAR berupa preposisi. Gambar 4 memperlihatkan pembagian PAR dalam bahasa Arab.



Gambar 3. Pembagian VERB



Gambar 4. Pembagian PAR

3. HASIL TERJEMAHAN

Baseline yang digunakan pada penelitian ini adalah hasil translasi dari Moses yang merupakan salah satu mesin translasi yang menggunakan metode statistik. Spesifikasi sistem yang digunakan sebagai baseline dalam penelitian adalah sebagai berikut:

ISSN Print : 2085-1588
ISSN Online : 2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index

1. Mesin translasi dilatih dengan GIZA++[4] menggunakan 6226 pasang kalimat berbahasa Arab – Indonesia yang diambil dari ayat-ayat Al-Quran beserta terjemahannya,
2. Model bahasa trigram dibangun dengan SRILM [5] menggunakan 6226 kalimat berbahasa Indonesia yang merupakan terjemah dari ayat-ayat Al-Quran.

Setelah dilakukan pengecekan terhadap kalimat-kalimat yang memiliki kata yang dapat mengandung beberapa makna, ditemukan kata-kata yang diterjemahkan tidak sesuai dengan kelas katanya sehingga menghasilkan hasil terjemahan yang kurang tepat.

Tabel 2. Kesalahan Penerjemahan Kata

S	b$r AlmnAfqyn b>n lhm E*AbA >lymA
T	seorang orang-orang munafik itu adalah karena sesungguhnya bagi mereka azab yang pedih
R	kabarkanlah kepada orang-orang munafik bahwa mereka akan mendapat siksaan yang pedih

S: Sumber; T: Hasil Translasi; R: Referensi

Kata yang terdiri dari susunan huruf ‘ba’, ‘syin’ dan ‘ro’ pada surat An-Nisaa: 138 (Tabel 2) seharusnya dibaca “basysyir” yang berarti “berilah kabar gembira” atau ”kabarkanlah” karena kedudukan katanya sebagai kata benda. Akan tetapi, mesin translasi justru mengartikannya dengan “seorang” yang memiliki kedudukan sebagai kata benda. Hal ini sangat mungkin terjadi karena mesin translasi ini menggunakan metode statistik untuk memilih frase yang bersesuaian dengan frase sumber tanpa melihat kedudukan kata tersebut dalam kalimat.

Kata b$r dapat menduduki beberapa kelas kata sehingga dapat memiliki beberapa makna. Tabel 3 menunjukkan pilihan hasil translasi dari kata “b$r”. Pilihan hasil translasi yang ada mengandung makna dari kata untuk setiap kelas kata.

Tabel 3. Hasil Translasi Kata "b$r"

1	b$r ||| berilah kabar gembira kepada
2	b$r ||| berilah kabar gembira
3	b$r ||| berilah kabar
4	b$r ||| berilah
5	b$r ||| gembirakanlah
6	b$r ||| bagi seseorang manusia
7	b$r ||| hanyalah seorang manusia
8	b$r ||| ini manusia
9	b$r ||| itu diajarkan oleh seorang manusia
10	b$r ||| itu diajarkan oleh seorang
11	b$r ||| manusia
12	b$r ||| oleh seorang
13	b$r ||| seorang manusia
14	b$r ||| seorang
15	b$r ||| seseorang manusia

ISSN Print : 2085-1588
ISSN Online : 2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index

4. PENANGANAN KELAS KATA

Kesalahan pemilihan makna disebabkan karena mesin translasi tidak mengenali kedudukan kata dalam kalimat. Permasalahan ini bisa diselesaikan dengan cara menambahkan informasi mengenai kedudukan suatu kata dalam kalimat pada dokumen sumber sebelum proses *training* dilakukan (*PoS tag*). Tabel 4 menunjukkan contoh kalimat yang telah ditambahkan *PoS tag*. Kata "AlHmd" diidentifikasi memiliki PoS tag "noun_prop" oleh sistem. Sehingga kata "AlHmd" ditulis dengan "AlHmd-noun_prop" pada dokumen yang baru.

Table 4. Dokumen Yang Diberi PoS Tag

No	Kalimat
1	AlHmd-noun_prop llh-noun_prop rb-noun AlEAlmyn-noun
2	AlrHmn-noun_prop AlrHym-noun_prop
3	AyAk-part nEbd-verb wAyAk-part nstEyn-verb
4	SrAT-noun Al*yn-pron_rel AnEmt-verb Elyhm-prep gyr-noun AlmgDwb-adj Elyhm-prep wlA-part_neg AlDAlyn-noun
5	wAl*yn-pron_rel y&mnwn-verb bmA-pron_rel Anzl-verb Alyk-prep wmA-pron_rel Anzl-verb mn-prep qblk-noun wbAlAxrp-noun hm-pron ywqnwn-verb

Dokumen berbahasa Arab yang sudah dimodifikasi ini yang nanti akan menjadi data latih mesin translasi yang baru. Sehingga ketika melakukan *training*, mesin translasi akan menangani kata-kata yang memiliki kedudukan yang berbeda dalam suatu kalimat dengan mengelompakkan kata-kata tersebut dalam kelompok yang berbeda.

PoS tagger yang digunakan dalam penelitian ini dilakukan secara manual. Penelitian ini kemudian menguji mesin translasi dengan 100 kalimat berbahasa Arab dengan PoS tag manual. Diantaranya terdapat 25 kalimat mengandung kata yang dapat memiliki beberapa kelas kata. Sebelumnya, mesin translasi dilatih dengan menggunakan 6226 pasang kalimat berbahasa Arab – Indonesia dengan PoS tag manual. Model bahasa dibangun dengan menggunakan 6226 kalimat berbahasa Indonesia dengan PoS tag manual. Tabel 5 menunjukkan contoh hasil dengan menggunakan *POS Tagger*.

Tabel 5. Hasil Tranlasi dengan *PoS Tagger*

Kalimat
B: seorang orang-orang munafik itu adalah karena sesungguhnya bagi mereka azab yang pedih
S: gembirakanlah orang-orang munafik itu adalah karena sesungguhnya mereka dengan azab yang pedih

B: Tanpa PoS tag; S: Dengan PoS tag

Hasilnya adalah tingkat efektifitas sebesar 20% atau sebanyak 5 dari 25 kalimat yang mengandung kata dengan beberapa kelas kata bisa diterjemahkan dengan baik sesuai

ISSN Print : 2085-1588
ISSN Online : 2355-4614
http://ejournal.unsri.ac.id/index.php/jsi/index

dengan kelas katanya menggunakan mesin penerjemah berbasis statistik yang telah mengimplementasi solusi yang diusulkan dalam penelitian ini berupa pemberian PoS tag.

5. KESIMPULAN DAN SARAN

Kata yang dapat memiliki beberapa kelas kata menjadi salah satu penyebab buruknya hasil translasi mesin penerjemah berbasis statistik dari bahasa Arab ke bahasa Indonesia. Hal ini dikarenakan mesin translasi berbasis statistik tidak melihat kedudukan kata tersebut dalam kalimat sehingga sangat mungkin terjadi pemilihan frase yang memiliki *PoS tag* yang tidak sesuai.

Salah satu solusi yang bisa dilakukan untuk mengurangi dampak ini adalah dengan cara menambahkan informasi mengenai kedudukan suatu kata dalam kalimat pada dokumen sumber sebelum proses *training* dilakukan berupa *PoS tag*.

Penelitian ini juga menyarankan agar dilakukan penelitian lebih lanjut untuk menemukan solusi yang lebih baik dalam menangani kasus dimana suatu kata dapat memiliki beberapa kelas kata yang mengakibatkan perbedaan makna.

DAFTAR PUSTAKA

[1] Hadj, Y.O. Mohamed El., Al-Sughayeir, I.A., Al-Ansari, A.M. 2009. Arabic Part-Of-Speech Tagging Using The Sentence Structure. Imam University
[2] Koehn, Philipp. 2007. Statistical Machine Translation. The University of Edinburgh.
[3] Och, Franz Josef., Ney, Hermann. 2000. Statistical Machine Translation. RWTH Aachen University.
[4] Och, Franz Josef., Ney, Hermann. 2003. A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, volume 29, number 1, pp. 19-51.
[5] SRILM - The SRI Language Modeling Toolkit. http://www.speech.sri.com/projects/srilm/. Waktu akses 28 Februari 2013.