KNSI2014-42

DAMPAK GABUNGAN KATA ARAB TERHADAP HASIL MESIN PENERJEMAH BERBASIS STATISTIK

Rahmat Izwan Heroza M.T.[1]

[1,2]Jurusan Sistem Informasi, Fakultas Ilmu Komputer, Universitas Sriwijaya
[3]Jl. Raya Palembang-Prabumulih KM 32 Indralaya Ogan Ilir. 30662
[1]rahmatheroza@unsri.ac.id

Abstrak

Tulisan ini meneliti dampak adanya gabungan kata (*muta 'allaq*) terhadap hasil mesin penerjemah berbasis statistik dalam menerjemahkan dokumen berbahasa Arab ke dalam bahasa Indonesia. Uniknya, kata-kata yang tergabung dalam *muta'allaq* tidak harus muncul secara berurutan. *Muta 'allaq* terkadang dipisahkan oleh banyak kata yang tidak terikat jumlah maupun jenisnya dengan *muta 'allaq*. Penelitian ini mengungkap bahwa hanya 2 dari 25 kalimat yang mengandung *muta 'allaq* berhasil diterjemahkan dengan baik. Penelitian ini akhirnya mengusulkan solusi untuk mengurangi dampak adanya *muta 'allaq* dengan cara pengidentifikasian *muta 'allaq*. Setelah teridentifikasi, kata -kata penyusun *muta 'allaq* ini diletakkan pada posisi yang berdekatan sehingga metode statistik dapat merekam kemunculan kedua gabungan kata ini. Hasilnya adalah tingkat efektifitas sebesar 60% atau sebanyak 15 dari 25 kalimat yang mengandung *muta'allaq* bisa diterjemahkan dengan baik menggunakan mesin penerjemah berbasis statistik yang telah mengimplementasi solusi yang diusulkan dalam penelitian ini. Solusi ini juga meningkatkan nilai BLUE Score dari semula 0.3574 menjadi 0.3584 atau meningkat sebesar 0.0010 poin.

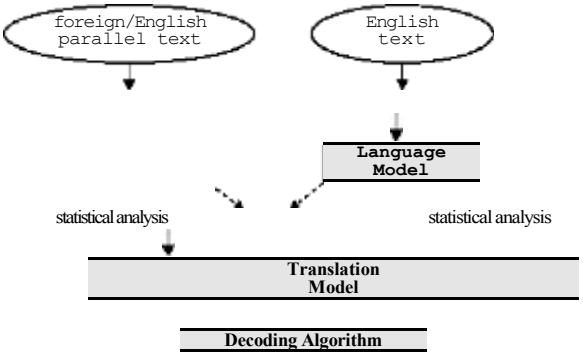
Kata kunci : mesin translasi berbasis statistik bahasa Arab, *muta 'allaq*, Moses, PoS tag, MADA+TOKAN

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 [4]. Sebuah mesin translasi berbasis statistik memiliki dua komponen utama yaitu training pipeline dan decoder. Training pipeline terdiri

Statstical machine translation

dari translation model dan language model (Gambar 1) [2].



Gambar 1. Komponen Mesin Translasi Berbasis Statistik

Penulis tertarik untuk meneliti dampak yang akan terjadi ketika metode statistik digunakan dalam menerjemahkan dokumen berbahasa Arab ke dalam bahasa Indonesia. Hal ini dikarenakan bahasa Arab memiliki gabunga n kata (*muta 'allaq*) yaitu suatu kata yang selalu diikuti oleh harfjar (preposisi) [3]. Uniknya, gabungan kata pada bahasa arab *(muta 'allaq)* memiliki karakteristik yang berbeda dengan gabungan kata pada bahasa Indonesia.

Baseline yang digunakan pada penelitian ini adalah hasil translasi dari Moses yang merupakan salah satu mesin translasi yang menggunakan metode statistik.

Di akhir tulisan, penelitian ini menguji salah satu usulan solusi yang bisa digunakan untuk mengatasi dampak yang muncul akibat adanya *muta 'allaq* dalam dokumen yang berupa buruknya hasil terjemahan mesin translasi berbasis statistik dari bahasa Arab ke bahasa Indonesia.

2. *Muta'allaq*

Muta 'allaq adalah suatu kata yang selalu diikuti oleh harf jar (preposisi) [3]. Kata pada *muta 'allaq* berupa kata kerja. Preposisi yang mengikuti kata *muta 'allaq* tidak sama untuk setiap kata muta'allaq. Perbedaan preposisi yang mengikuti

suatu kata *muta'allaq* dapat memberikan makna yang berbeda pula. Bahkan satu makna dari *muta 'allaq* yang diikuti oleh suatu preposisi tertentu memiliki makna yang berlawanan jika muta'allaq tersebut diikuti oleh preposisi yang lain. Sebagai contoh kata "rgb fy" (roghiba fii) memiliki makna"mencintai", sedangkan kata "rgb En" (roghiba Ôan) memiliki makna "membenci".

Adapun kata-kata yang dikelompokkan ke dalam preposisi adalah:

1. (ba)
2. sJi (ila)
3. (ka)
4. J(la)
5. عن
6. من
7. '-(fi)
8. على(Ôa

Ada tiga kategori perilaku mutaÕallaq dalam bahasa Arab. Yang pertama, mutaÕallaq selalu diikuti oleh preposisi. Kata ini tidak mempunyai makna jika tidak diikuti oleh preposisi. Preposisi yang mengikutinya mempunyai implikasi terhadap makna yang dikandungnya. Kedua, mutaÕallaq yang mempunyai makna tertentu jika diikuti oleh preposisi, akan tetapi tidak bermakna jika tidak diikuti oleh preposisi terebut. Ketiga, mutaÕallaq yang diikuti atau tidak oleh preposisi, maknanya tetap sama.

3. Identifikasi *Muta'allaq*

Setelah dilakukan pengecekan terhadap 25 kalimat yang memiliki mutaÕallaq, hanya 2 kalimat yang berhasil teridentifikasi memiliki mutaÕallaq sehingga memberikan hasil terjemah yang tepat.

Tabel 1. Kesalahan Penerjemahan *Muta 'alla*

S	vm ytwb Allh mn bEd *lk ElY mn y$A' wAllh gfwr rHym
T	kemudian mereka bertaubat Allah sesudah itu kepada siapa yang dikehendaki-nya dan Allah maha pengampun lagi maha
R	sesudah itu Allah menerima taubat dari orangorang yang dikehendaki-nya Allah maha pengampun lagi maha penyayang

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

Mesin translasi berbasis statistik ternyata kesulitan dalam mengidentifikasi mutaÕallaq yang terdiri dari dua kata, yang mana kedua kata ini dipisahkan oleh banyak kata. Salah satu contoh kalimat yang mengandung *muta'allaq* yang tidak bisa diterjemahkan dengan baik oleh mesin penerjemah berbasis statistik terdapat pada surat AtTaubah: 27 (Tabel 1).

Mesin translasi berbasis statistik memiliki kemampuan untuk mengingat frase yang tersusun dari hingga tujuh buah kata. Kata-kata ini dianggap mungkin memiliki makna yang khusus apabila sering muncul secara berurutan. Ini lah yang digunakan mesin translasi berbasis statistik sebagai indikasi dari gabungan kata, dimana kata yang berdiri sendiri mungkin berbeda maknanya apabila kata tersebut diikuti oleh kata lain. Seperti pada *muta 'allaq*, kata kerja yang berdiri se ndiri berbeda maknanya apabila kata kerja tersebut memiliki preposisi sehingga menjadi sebuah *muta 'allaq*.

Akan tetapi, kata-kata yang tergabung dalam mutaÕallaq tidak harus muncul secara berurutan.

MutaÕallaq terkadang dipisahkan oleh banyak kata yang tidak terikat jumlah maupun jenisnya dengan mutaÕallaq. Dan tidak mungkin untuk merekam semua kata-kata yang memisahkan mutaÕallaq dan memasukkannya ke dalam table frase. Hal ini lah yang belum bisa ditangani oleh mesin penerjemah berbasis statistik. Kita dapat melihat kesalahan identifikasi mutaÕallaq yang berupa gabungan kata "ytwb" (yatubu) + "ElY" (Õala) yang berarti "menerima taubat". Mesin translasi berbasis statistik menerjemahkannya menjadi "bertaubat" yang merupakan terjemah dari kata "ytwb" (yatubu) (Tabel 2).

Tabel 2. Kesalahan Identifikasi *Muta 'alla*

S	vm ytwb Allh mn bEd *lk ElY mn y$A' wAllh gfwr rHym
T	[[0..0]:kemudian mereka] [[1..1]:bertaubat] [[2..2]:allah] [[3..5]:sesudah itu] [[6..7]:kepada siapa yang] [[8..9]:dikehe ndaki -nya dan allah] [[10..10]:maha pengampun] [[1 1..1 1]:lagi maha penyayang]

S: Sumber; T: Hasil

Translasi **4. Penanganan *Muta***

'allaq

Karena penulisan *muta 'allaq* tidak harus berurutan, sebuah proses harus melakukan pengecekan terhadap ada atau tidaknya preposisi yang mengikuti sebuah kata kerja. Oleh karena itu, diperlukan sebuah proses untuk mengenali kedudukan sebuah kata apakah termasuk kata kerja atau termasuk preposisi. Hal ini bisa dilakukan dengan menggunakan PoS tag. Lalu apabila diketahui bahwa terdapat kata kerja yang kemudian diikuti oleh preposisi pada kata-kata setelahnya dalam kelimat, maka dilakukan penggeseran preposisi sehingga berada tepat setelah kata kerja. Dengan cara ini, gabungan kata yang merupakan *muta 'allaq* dapat disimpan pada tabel frase.

PoS tagger yang digunakan dalam penelitian ini adalah MADA+TOKAN yang dikembangkan oleh peneliti di Center for Computational Learning Systems (CCLS)

Universitas Columbia [1]. MADA+TOKAN adalah sebuah sistem pemrosesan bahasa alami (NLP) yang memiliki fungsi *token ization, diacrization, morphological disambiguation, PoS tagging, stemming* dan *lemmatization* untuk bahasa Arab.

Penelitian ini kemudian menguji mesin translasi dengan 100 kalimat berbahasa Arab dengan PoS tag hasil dari MADA. Diantaranya terdapat 25 kalimat mengandung *muta 'allaq* yang tidak bisa diidentifikasi dengan baik oleh mesin penerjemah berbasis statistik. Sebelumnya, mesin translasi dilatih dengan menggunakan 6226 pasang kalimat berbahasa Arab – Indonesia dengan PoS tag hasil dari MADA. Model bahasa dibangun dengan menggunakan 6226 kalimat berbahasa Indonesia dengan PoS tag hasil dari MADA.

Hasilnya adalah tingkat efektifitas sebesar 60% atau sebanyak 15 dari 25 kalimat yang mengandung *muta 'allaq* bisa diterjemahkan dengan baik menggunakan mesin penerjemah berbasis statistik yang telah mengimplementasi solusi yang diusulkan dalam penelitian ini. Solusi ini juga meningkatkan nilai BLUE Score dari semula 0.3574 menjadi 0.3584 atau meningkat sebesar 0.0010 poin.

5. Kesimpulan dan Saran

Gabungan kata pada bahasa Arab atau *muta 'allaq* menjadi salah satu penyebab buruknya hasil translasi mesin penerjemah berbasis statistik dari bahasa Arab ke bahasa Indonesia. Hal ini dikarenakan kata-kata yang tergabung dalam muta'allaq tidak harus muncul secara berurutan. Muta'allaq terkadang dipisahkan oleh banyak kata yang tidak terikat jumlah maupun jenisnya dengan muta'allaq. Dan tidak mungkin untuk merekam semua kata-kata yang memisahkan muta'allaq dan memasukkannya ke dalam table frase.

Salah satu solusi yang bisa dilakukan untuk mengurangi dampak ini adalah dengan cara mengidentifikasi terlebih dahulu kemunculan *muta 'allaq*. Hal ini bisa dilakukan dengan menggunakan PoS tagger untuk mengidentifikasi *muta 'allaq* yang terdiri dari kata kerja dan preposisi. Setelah diidentifikasi kemunculan kata kerja dan preposisi dalam satu kalimat, preposisi kemudian digeser sehingga berada tepat setelah kata kerja.

Penelitian ini juga menyarankan agar dilakukan penelitian lebih lanjut untuk menemukan solusi dalam mengidentifikasi *muta 'allaq* dan menangani kasus dimana suatu kalimat mengandung *muta 'allaq*..

Daftar Pustaka:

[1] Habash, Nizar., Rambow, Owen., Roth, Ryan. *MADA+TOKAN Software Suite*. http://www1.ccls.columbia.edu/MADA/. Center for Computational Learning Systems (CCLS), Columbia University. Waktu akses 1 Februari 2013.

[2] Koehn, Philipp. 2007. *Statistical Machine Translation*. The University of Edinburgh.

[3] Nurabayan, Yayan. 2007. *Peningkatan Kemampuan Mahasiswa dalam Menulis Skripsi melalui Pengenalan Muta 'allaq*. Universitas Pendidikan Indonesia.

[4] Och, Franz Josef., Ney, Hermann. 2000. *Statistical Machine Translation*. RWTH Aachen University.