Fitur-fitur morfologi untuk analisis perkataan bahasa Melayu

Mohd Yunus Sharum
m_yunus@ upm.edu.my

Kata kunci: Pemprosesan bahasa tabii, morfologi komputasi, analisis morfologi bahasa Melayu.

Pengenalan

Ketaksaan dalam bahasa wujud dalam beberapa aspek merangkumi ketaksaan leksikal akibat *homograf* dan *homonim* (sama ejaan), atau ketaksaan makna yang wujud pada *polisem* (perkataan dengan lebih daripada satu makna). Masalah ketaksaan menyukarkan perisian komputer melakukan proses analisis perkataan bahasa Melayu, dan mendorong kepada penghasilan output atau hasil analisis yang tidak tepat. Misalnya ketaksaan leksikal akibat homonim menyukarkan analisis untuk mengenal pasti kata akar bagi suatu terbitan. Antaranya termasuklah kata terbitan seperti ‘mengecam’ (cam / kecam) dan ‘mengapit’ (apit / kapit). Dalam hal ini, perisian bermasalah untuk menentukan kata akar yang perlu dihasilkan kerana terdapat lebih daripada satu kemungkinan output bagi kata terbitan tersebut.

Masalah ketaksaan boleh diatasi dengan menggunakan beberapa kaedah. Salah satunya ialah dengan mengenal pasti makna perkataan tersebut. Misalnya ‘mengecam’ yang bersinonim dengan perkataan *mengutuk* dikenal pasti menggunakan kata akar ‘kecam’, manakala ‘mengecam’ yang bersinonim dengan *mengenal pasti* akan menghasilkan kata akar ‘cam’. Kaedah lain ialah berdasarkan konteks dan konkordans, iaitu berdasarkan perkataan-perkataan lain yang bersebelahan atau berhampiran dengannya. Misalnya ‘... mengecam perbuatan ...’ menunjukkan ‘mengecam’ yang bersumberkan ‘kecam’ manakala ‘... mengecam wajah ...’ menunjukkan ‘mengecam’ yang bersumberkan kata akar ‘cam’. Namun dalam kebanyakan pelaksanaan, maklumat seperti makna, konteks dan konkordans tidak boleh didapati dengan mudah, apatah lagi jika perisian itu hanya diberi input dalam bentuk perkataan tunggal (tiada maklumat makna atau konteks penggunaannya).

Fitur morfologi boleh membantu proses analisis perkataan yang bersandarkan kepada input berbentuk perkataan tunggal kerana fitur morfologi terkandung dalam perkataan dan bukannya persekitaran. (Analoginya, seumpama kita mengenal pasti struktur suatu sampel otak berdasarkan pengetahuan neurologi tanpa mengetahui otak tersebut daripada monyet, tikus mahupun haiwan lain.) Fitur morfologi didefinasikan oleh penulis sebagai ciri-ciri yang dipamerkan atau dipunyai oleh struktur morfologi dalam suatu perkataan. Misalnya kata bahasa Melayu tidak mempunyai lebih daripada tiga lapis imbuhan awalan (Hassan, 1974; 1988). Sebahagian fitur morfologi terhad untuk bahasa tertentu, seperti *kekangan tiga imbuhan awalan* tersebut, yang khusus untuk bahasa Melayu. Namun ada fitur morfologi yang dimiliki oleh lebih daripada satu bahasa, misalnya *suatu kata akar boleh digabungkan dengan imbuhan akhiran* (bahasa Melayu, Inggeris, Jerman dll).

Amnya bahasa Melayu mengandungi beberapa fitur morfologi yang sering dipertimbangkan dalam pelaksanaan analisis morfologi. Antaranya morfotaktik (Bali, 2004; Sulaiman et. al, 2011) dan sifat khusus klitik dan partikel (Bali, 2004; Asian et. al, 2005; Adriani et. al, 2007). Namun begitu fitur-fitur ini belum pernah dikaji secara khusus untuk mengenal pasti kesannya terhadap hasil proses analisis perkataan, khususnya untuk bahasa Melayu. Makalah ini membincangkan tiga fitur morfologi bahasa Melayu untuk mengatasi masalah ketaksaan

leksikal dalam mengenal pasti kata akar suatu kata terbitan, sekaligus meningkatkan ketepatan hasil analisis morfologi yang dijalankan terhadap perkataan bahasa Melayu dengan hanya bersumberkan maklumat di peringkat morfologi.

Proses Analisis Morfologi

Proses analisis morfologi ialah proses untuk merungkai dan mendapatkan maklumat-maklumat morfologi yang terkandung dalam suatu perkataan. Misalnya contoh berikut memaparkan penghasilan kata akar ‘makan’ daripada kata terbitan ‘pemakanan’:

pemakanan => pe+**makan**+an => **makan**

Terdapat beberapa teknik yang diterapkan untuk melaksanakan analisis morfologi, merangkumi pendekatan heuristik dan statistik, serta penggunaan leksikon. Penulis menggunakan teknik leksikon (suatu set perkataan), di mana suatu leksikon yang mengandungi set kata akar digunakan untuk mengenal pasti sama ada ‘kata akar’ yang hendak dihasilkan sebagai output adalah sah ataupun tidak selepas dilakukan proses membuang imbuhan. Misalnya dalam contoh berikut,

pemakanan => pe+**makan**+an => **makan**

berkat => ber+**kat** => **kat***
=> **berkat**

Hasil analisis ‘makan’ ialah kata akar yang sah kerana kata sedemikian wujud dan disimpan dalam leksikon. Maka ia boleh dihasilkan sebagai output. Tetapi ‘kat’ tidak wujud dan ia tiada dalam leksikon. Maka ‘kat’ bukan kata akar yang sah. Bagi input ‘berkat’ akan menghasilkan kata akar ‘berkat’ juga dan ia akan dihasilkan sebagai output.

Proses analisis morfologi banyak dimanfaatkan dalam aplikasi komputer yang melibatkan pemprosesan bahasa tabii. Antara aplikasi masakini ialah proses mendapatkan kata akar suatu perkataan yang digunakan dalam capaian semula maklumat, serta proses mengetag korpus (mengenal pasti kata akar, yang kemudian menggunakan maklumat tersebut untuk mengenal pasti kelas kata).

Fitur-Fitur Morfologi dalam Kata Bahasa Melayu

Penulis menggunakan tiga fitur morfologi dalam bahasa Melayu untuk membantu mengurangkan ketaksaan dan meningkatkan ketepatan analisis morfologi oleh perisian. Fitur-fitur tersebut ialah sifat unik perkataan ekasuku kata, morfotaktik dalam pengimbuhan, dan sifat khusus klitik dan partikel.

Secara amnya kata terbitan bagi perkataan ekasuku kata bahasa Melayu dapat dibezakan daripada kata terbitan yang bersumberkan perkataan berbilang suku kata. Penerbitan kata daripada perkataan ekasuku kata yang melibatkan imbuhan awalan meN- dan peN- akan menerbitkan kata terbitan yang mempunyai imbuhan awalan *menge-* dan *penge-*. Manakala bagi imbuhan lain seperti beR- (*ber-*) dan teR- (*ter-*), akan kekal sebagai imbuhan yang asal. Berdasarkan fitur ini, maka suatu set perkataan ekasuku kata (yang menggunakan imbuhan *menge-* dan *penge-*) boleh dibina dan dipisahkan daripada set perkataan berbilang suku kata.

Kaedah ini boleh membantu mengatasi ‘ketaksaan leksikal’ pada kata terbitan ‘mengepit’ (kepit/pit). Amnya ‘pengepit’ dan ‘mengepit’ bagi ‘pit’ belum wujud. Tetapi ‘berpit’ boleh wujud. Jika pemisahan leksikon tidak dibuat, ‘pengepit’, ‘mengepit’ dan ‘berpit’ sukar dikawal dan kesemuanya akan menghasilkan output ‘pit’ sebagai kata akar. Melalui pemisahan leksikon, maka ‘pit’ boleh disimpan sebagai berbilang ‘suku kata’ supaya mengelakkan ‘mengepit’ dan ‘pengepit’ disalahtafsir sebagai ‘meN+pit’ dan ‘peN+pit’ (menghasilkan kata akar yang tidak sah). Tetapi struktur ‘beR+pit’ adalah sah kerana imbuhan ‘beR’ melibatkan pencarian dalam kedua-dua leksikon (maka ‘pit’ output yang sah bagi ‘berpit’).

Fitur kedua ialah morfotaktik pengimbuhan bahasa Melayu. Amnya pengimbuhan bahasa Melayu tertakluk kepada set peraturan pengimbuhan yang terhad (finit). Berdasarkan analisis penulis, pengimbuhan bahasa Melayu terhad kepada 78 kombinasi imbuhan (dengan imbuhan –i dan –kan disatukan sebagai –i / –kan). Antara peraturan-peraturan imbuhan tersebut ialah ber+(pe+_+an) (*berpekerjaan*), ber+(peN+_+an) (*berpendapatan*), ber+(peR+_+an) (*berperaturan*), ber+(peN+(ke+_+an)) (*berpengetahuan*) dll. Morfotaktik dapat membantu mengatasi masalah ketaksaan leksikal yang melibatkan suku kata berupa imbuhan, misalnya ketaksaan melibatkan analisis ke atas kata terbitan ‘kediaman’. Dalam contoh ini, ‘di...’ boleh disalahtafsir sebagai imbuhan. Tetapi masalah ini dapat dielakkan melalui tapisan morfotaktik, kerana kombinasi ke+di+_ bukan merupakan kombinasi yang sah.

Fitur ketiga ialah sifat khusus klitik dan partikel dalam bahasa Melayu seperti ku-, kau-, -lah, -tah dll. Klitik dan partikel amnya boleh dianggap imbuhan kerana sifatnya yang seakan imbuhan. Namun klitik dan partikel tidak bertindak sepertimana imbuhan-imbuhan terbitan peN-, meN-, dll. kerana klitik dan partikel mempunyai kedudukan yang khusus dalam struktur kata. Amnya, klitik dan partikel tidak akan merapati atau bersambung dengan kata akar sekiranya wujud imbuhan terbitan pada kata akar tersebut. Maka hal ini mewujudkan syarat dan kekangan yang membezakan antara imbuhan dengan yang bukan imbuhan (suku kata). Misalnya ‘...lah’ dalam ‘permasalahan’ boleh dibezakan sebagai suku kata kerana selepasnya wujud imbuhan akhiran -an. Tanpa kekangan kedudukan klitik dan partikel, ‘lah’ mungkin akan disalahtafsir sebagai partikel.

Dapatan Kajian

Ketiga-tiga fitur morfologi tersebut telah diimplimentasi sebagai fungsi-fungi yang digunakan dalam sebuah perisian penganalisis morfologi. Penganalisis morfologi tersebut digunakan untuk menganalisis set 14319 perkataan yang perolehi daripada masukan dalam buku Tesaurus Bahasa Melayu Dewan (DBP, 2005).

Daripada eksperimen tersebut, didapati teknik implimentasi fitur berkaitan ciri khusus perkataan ekasuku kata telah dapat meningkatkan 0.67% ketepatan analisis, serta mengurangkan 0.64% kesilapan akibat ketaksaan. Manakala bagi fitur berkaitan morfotaktik pula telah meningkatkan 1.55% ketepatan analisis selain mengurangkan 1.53% kesilapan akibat ketaksaan. Fitur melibatkan sifat khusus klitik dan partikel pula telah meningkatkan ketepatan analisis sebanyak 0.39%. Walaupun fitur ini tidak mengurangkan kesilapan daripada masalah ketaksaan, namun penerapan fitur ini untuk analisis telah meningkatkan jumlah kepersisan analisis (misalnya untuk 3 hasil output bagi 1 input, 2:3 adalah hasil yang tepat, berbanding 1:3 dalam pencapaian sebelumnya).

Kesimpulan

Berdasarkan dapatan kajian, didapati fitur-fitur morfologi amnya mempunyai kesan positif terhadap analisis morfologi. Ini membuktikan bahawa maklumat-maklumat yang terkandung pada peringkat morfologi saja pun boleh memberikan impak kepada proses analisis morfologi, tanpa melibatkan maklumat dari luar peringkat morfologi seperti konteks perkataan, konkordans dan sebagainya. Ini memungkinkan analisis morfologi dilakukan dengan berdasarkan input tunggal (input berbentuk perkataan).

Namun begitu, impak fitur-fitur morfologi terhadap hasil analisis morfologi agak kecil. Hal ini mungkin disebabkan maklumat-maklumat yang terhasil melalui penggunaan fitur bersifat khusus dan terpencil. Selain itu impak terhadap penyahtaksaan juga kecil kerana peratusan ketaksaan dalam bahasa sememangnya kecil. Namun ini tidak bermakna perubahan ini tidak signifikan kerana ketaksaan leksikal dalam suatu bahasa boleh berubah mengikut perkembangan leksikon bahasa tersebut. Bahkan hasil eksperimen menunjukkan perubahan terhadap ketepatan analisis yang melibatkan ketaksaan, melibatkan lebih dari 100 perkataan.

Pada masa depan, kajian lebih mendalam akan dilakukan bagi membandingkan fitur-fitur dan mengenal pasti fitur manakah yang memberikan impak tertinggi kepada proses analisis. Selain itu, penulis cuba mengenal pasti fitur-fitur lain, yang mungkin wujud dan memberi impak yang lebih besar berbanding fitur-fitur yang telah diuji.

Rujukan

Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M., and Williams, H.E. (2007). Stemming Indonesian: A Confix-stripping Approach. *ACM Transactions on Asian Language Information Processing (TALIP)*. 6(4). Dec. 2007:1-33.

Asian, J., Williams, H.E., and Tahaghoghi, S.M. Stemming Indonesian. In Proceedings of the Twenty-Eighth Australasian Conference on Computer Science – Volume 38 (Newcastle, Australia). V. Estivill-Castro Ed.; ACSC, vol. 102. Australian Computer Society: Darlinghurst, Australia, 2005.

Bali, R.M. *Computational Analysis of Affixed Words in Malay Language*. Paper presented in the 8th International Symposium on Malay/Indonesian Linguistics (ISMIL8), Penang, Malaysia, 2004.

DBP (2005). *Tesaurus Bahasa Melayu Dewan – Edisi Baharu*. Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Hassan, A. (1974). *The Morphology of Malay*. Kuala Lumpur, Malaysia: Dewan Bahasa dan Pustaka.

Hassan, A. (1988). *Penerbitan Kata Dalam Bahasa Malaysia*. PJ, Selangor, Malaysia: Penerbit Fajar Bakti Sdn. Bhd.

Sulaiman, S., Gasser, M., and Kübler, S. Towards A Malay Derivational Lexicon: Learning Affixes using Expectational Maximization. In Proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), IJCNLP, 2011.