STEMMING WORDS DENGAN N-GRAM DAN LEXEME BASED UNTUK TEKS BERBAHASA KOREA

STEMMING WORDS WITH N-GRAM AND LEXEME BASED FOR KOREAN LANGUAGE TEXT

Nadya Eka Putri P.[1], Shaufiah[2], Mira Kania S.[3]

[1,2,3]Prodi S1 Teknik Informatika, Fakultas Informatika, Universitas Telkom
[1]nadya.permadi@gmail.com, [2]shaufiah@telkomuniversity.ac.id, [3]mirakania@telkomuniversity.ac.id

Abstrak

Bahasa Korea termasuk ke dalam bahasa aglutinatif yang termasuk unik dan memiliki berbagai jenis pelekatan morfem, dengan kondisi ini, pengaplikasian teknik stemming words dianggap sedikit sulit untuk dilakukan. Beberapa penelitian sudah dilakukan, namun masih ditemui beberapa kesalahan dikarenakan adanya keunikan dari karakter kata dalam Bahasa Korea. Dalam penelitian kali ini akan dibahas teknik baru untuk melakukan stemming words atau pencarian kata dasar disertai dengan deteksi imbuhannya. Penelitian ini bertujuan untuk membentuk kata dasar dari kata kerja berimbuhan pada bahasa Korea dan mencari jenis dan arti dari imbuhan yang melekat pada kata tersebut. Penelitian ini dilakukan dengan menggabungkan metode N-gram dan Lexeme Based. Dalam pencarian kata dasar ini sejumlah kata kerja yang mendapat imbuhan dalam tata bahasa tertentu dipecah untuk menghasilkan kata dasar dan imbuhan yang sesuai. Pemecahan kata berimbuhan dilakukan dengan metode N- gram dan dilanjutkan dengan pengaplikasian metode Lexeme Based untuk pencarian kata dasar serta jenis dan arti imbuhan. Hasil yang didapatkan pada penelitian ini adalah pembentukan kata dasar dan imbuhan yang disertai dengan jenis imbuhan serta arti dari imbuhan tersebut.

Kata kunci : stemming words, Bahasa Korea, N-gram, Lexeme Based, aglutinatif.

Abstract

Korean Language is known as an agglutinative language which is unique and has lot of morphems in it, so that it is a bit hard to apply the stemming words method in the usual way. Some researches has been done, yet there were still a few mistakes found. Now, this research talks about the new method of stemming words which able to proceed and form compound words into their own basic word and also detect the suffix or affix they use in it. The main goal from this research is to form a basic word from their compound words and detect the suffix or affix on it, includes the real meaning of the word and the meaning of the suffix or affix. This research proposed to combine the N-Gram method and the Lexeme Based method. To form a basic word, the compound words will be proceeded by dividing each syllables for basic word and suffix or affix. The dividing process will be done with N-Gram method, then continued with Lexeme Based method to start forming the basic word and the suffix or affix, includes finding their meanings. The final result from this research is to form the basic-word and the affixes, including the meaning of them.

Keywords : stemming words, Korean Language, N-Gram, Lexeme Based, agglutinative

1. PENDAHULUAN

1.1 Latar Belakang

Saat ini budaya mengenai Korea Selatan sedang sangat digemari oleh banyak orang di Indonesia dan bahkan di negara lainnya. Dimulai dari musik, sejarah, budaya, keindahan negara, hingga bahasanya yang sedang menjadi tren selama beberapa tahun belakangan ini. Banyak orang Indonesia saat ini bahkan turut mempelajari bahasa Korea untuk dapat mengikuti tren yang ada, namun dikarenakan struktur penggunaan bahasanya yang berbeda, mempelajari bahasa Korea dianggap lebih sulit bagi kebanyakan orang.

Bahasa Korea termasuk ke dalam bahasa aglutinatif, yaitu bahasa yang terbentuk atas pelekatan dari beberapa morfem dan partikel kata (Chaer, 2007*)*. Dalam penggunaannya, bahasa Korea memiliki banyak bentuk imbuhan untuk mendeskripsikan jenis tata bahasa yang berbeda dengan makna yang beragam. Dengan adanya keragaman ini, dibutuhkan suatu metode khusus yang dapat membantu untuk mempelajari bahasa Korea dalam menentukan makna kata yang sebenarnya dan untuk mempelajari penggunaan tata bahasa dari kata dasar tersebut. Salah satu teknik yang dapat digunakan yaitu teknik *stemming words*. Penelitian mengenai *stemming words* untuk bahasa aglutinatif seperti bahasa Korea sudah beberapa kali dilakukan, diantaranya untuk melakukan *indexing* dan peringkasan teks (Lee, 1996; Kim, 1996; Lee, 1999), namun masih ditemukan beberapa kesulitan dalam implementasi dan hasil yang kurang sesuai, dikarenakan oleh karakteristik huruf yang berbeda-beda dalam bahasa Korea dan penggunaan imbuhan yang beragam.

Pada penelitian kali ini teknik yang digunakan adalah penggabungan dari metode N-gram dan Lexeme Based. Melalui metode N-gram kata kerja berimbuhan dipecah menjadi beberapa bentuk potongan karakter string dan setelah itu diterapkan metode Lexeme Based dengan beberapa aturan tertentu sehingga bisa didapatkan hasil akhir berupa kata dasar dan makna dari kata tersebut, serta jenis tata bahasa dan imbuhan apa yang digunakannya. Dengan menggunakan teknik ini maka pencarian kata dasar dapat dilakukan melalui pemotongan string lalu dilanjutkan dengan pencarian makna sebenarnya yang terkandung dari kata tersebut, kemudian juga dapat diketahui jenis tata bahasa dan imbuhan yang digunakan, sehingga orang-orang dapat dengan mudah mempelajari dan memahami bahasa Korea. Dengan menggabungkan metode N-Gram dan Lexeme Based ini maka diharapkan penelitian pada *stemming words* untuk bahasa aglutinatif seperti Bahasa Korea dapat menjadi lebih baik.

2. LANDASAN TEORI

2.1 Morfologi Bahasa Korea

Bahasa Korea merupakan bahasa yang termasuk ke dalam rumpun bahasa Ural-Altai, seperti bahasa Jepang, Turki, Mongol, dan Tungu (Handbook of Korea, 1978*;* Chaer, 2007*).* Dengan perbedaan jenis tersebut maka bahasa Korea memiliki struktur tata bahasa dengan morfem khususnya sendiri. Pendapat yang sama juga dikatakan oleh Lee (2004), bahwa sistem tata bahasa yang digunakan dalam bahasa Korea adalah bidang kajian yang berkaitan dengan sistem pembentukan kata dari komponen morfem dan sistem pembentukan kalimat dari kata-kata (sintaksis). Bahasa Korea termasuk ke

dalam bahasa aglutinatif dimana kata-kata yang membentuknya terdiri dari gugus morfem. Bahasa aglutinatif sendiri merupakan bahasa yang memiliki kesatuan kata-kata tertentu sehingga dapat membentuk suatu kata baru dengan makna yang berbeda. Pelekatan atau kesatuan dalam bahasa aglutinatif dapat dengan mudah berubah maknanya saat satu atau dua kata dihilangkan dari kata tersebut, hal inilah yang menjadi fokus utama dalam morfologi bahasa Korea sebagai bahasa aglutinatif. Hal ini didukung oleh Choi et al (2009:67) yang juga menyatakan bahwa bahasa Korea sebagai bahasa aglutinatif memiliki pembentukan kalimat dengan cara menggabungkan morfem gramatikal pada suatu morfem leksikal. Untuk memperjelas, dapat diperhatikan contoh berikut :

(1) 수고하셨어요 [sugohasyôssôyo]
➔➔수고 [sugo] + 하 [ha] + 시 [si] + 었 [ôss] + 어요 [ôyo]
'Kerja yang bagus'

Contoh (1) merupakan jenis kalimat yang berifat lampau dan memiliki makna 'Kerja yang bagus'. Namun kalimat tersebut sebenarnya terdiri atas beberapa morfem terikat, diantaranya adalah nomina 수고 [sugo] yang bermakna 'kerja bagus', kemudian morfem yang melekat adalah 하 [ha] yang berarti komponen penjelas verba (V) untuk mengubah kata sebelumnya yang berbentuk nomina menjadi verba, kemudian morfem berikutnya adalah 시 [si] yang memiliki fungsi sebagai penanda honorifik, diikuti dengan 었 [ôss] yang memiliki peran sebagai pembentuk kata kerja lampau bagi kalimat, terakhir terdapat 어요 [ôyo] yang menjadi morfem penanda akhiran bagi kalimat tersebut.

Dari contoh (1) dapat juga diketahui bahwa dengan kalimat seperti itu di dalam bahasa Korea terdapat berbagai jenis imbuhan dan morfem yang melekat, sehingga dapat membentuk satu makna baru.

Kata dalam bahasa Korea juga dapat dikategorikan ke dalam 9 kategori yaitu nomina (명사), pronominal (대명사), numeralia (수사), verba (동사), adjektiva (형용사), artikel (관형사), adverbial (부사), interjeksi (감탄사) dan partikel (조사). Pengkategorian ini terbentuk beradasarkan perubahan morfem, fungsi, dan maknanya (Lee, 2006; Choi *et al*, 2009; Lee, 2007).

Kategori verba dalam bahasa Korea merupakan jenis kata kerja dasar yang tidak bisa berdiri sendiri sehingga selalu diperlukan pelekatan dari morfem lain dalam penggunaannya. Hal ini yang menjadi dasar pemahaman proses pelekatan morfem atau imbuhan di dalam kategori verba. Untuk pemahaman lebih lanjut dapat dilihat contoh berikut ;

(2) 밥을먹었어요[bab-eul môgôssôyo]
➔➔밥을 [bab-eul] + 먹 [môg] + 었 [ôss] + 어요 [ôyo]
'Sudah makan nasi'

Contoh (2) memiliki kata kerja dasar 먹 [môg] yang berarti 'makan', sedangkan untuk objeknya 밥 [bab] yang bermakna 'nasi' diikuti dengan morfem penjelas objek 을 [eul]. Untuk kata kerja-

nya sendiri, 먹 [môg] perlu diikuti dengan morfem lain yaitu 었 [ôss] yang menjelaskan penanda lampau dan 어요 [ôyo] sebagai penanda akhiran kalimat. Kata 먹 [môg] sendiri disini sebenarnya terbentuk dari kata kerja dasar 먹다 [môgda] yang berarti 'makan', namun kata kerja dasar tersebut harus dilekatkan pada morfem lain agar dapat digunakan di dalam kalimat yang baik. Oleh karena itu jenis kata kerja dasar dalam bahasa Korea selalu membutuhkan morfem dan partikel tambahan dalam penggunaannya.

2.2 Teknik Stemming Words

Stemming words merupakan suatu proses pemecahan kata-kata berdasarkan algoritma dan aturan yang ada. Biasanya proses *stemming words* digunakan sebagai suatu proses pencarian informasi atau *information retrieval* di dalam teks, kalimat dan indeks kata. Di dalam proses *stemming words* terdapat berbagai jenis metode dan algoritma yang dapat digunakan sebagai media pemrosesannya, diantaranya adalah N-gram, Lexeme Based, Morpheme Based, Algoritma Porter, Word-Based Indexing, dan masih banyak lagi. Dalam penelitian kali ini, yang berhubungan dengan bahasa Korea sebagai bahasa aglutinatif, maka metode yang digunakan adalah kombinasi metode N-Gram dengan Lexeme Based.

2.2.1 Metode N-Gram

N-gram adalah sebuah metode pemotongan atau pemisahan string di dalam kalimat atau kata. N-gram diaplikasikan untuk mengambil potongan string atau karakter dengan jumlah (n) tertentu dari sebuah kata yang ada secara berkelanjutan dari awal hingga akhir.

Metode N-gram dibedakan berdasarkan jumlah pemotongan string atau karakter yang diproses. Terdapat banyak jenis N-gram yaitu Uni-gram untuk pemotongan tiap satu karakter string, Bi-gram untuk pemotongan tiap dua karakter string, Tri-gram untuk pemotongan tiap tiga karakter string, Quad-gram untuk pemotongan tiap empat karakter string dan seterusnya

Contoh :

Teks : STEMMINGWORDS
Uni-gram : S, T, E, M, M, I, N, G, W, O, R, D, S
Bi-gram : _S, ST, TE, EM, MM, MI, IN, NG, GW, WO, OR, RD, DS, S_
Tri-gram : _ST, STE, TEM, EMM, MMI, MIN, ING, NGW, GWO, WOR, ORD, RDS, DS_
Quad-gram : _STE, STEM, TEMM, EMMI, MMIN, MING, INGW, NGWO, GWOR, WORD, ORDS, RDS_

Salah satu keunggulan dari penggunaan N-gram adalah metode ini tidak terlalu sensitif dengan kesalahan penulisan yang terdapat dalam suatu dokumen atau teks. Metode N-gram juga dinilai lebih efektif untuk digunakan dalam *stemming words* terhadap bahasa aglutinatif karena proses pemotongan karakter string yang dilakukan tidak bergantung pada jenis huruf dan bentuk huruf yang digunakan dalam bahasa tersebut.

2.2.2 Metode Lexeme Based

Dalam teori pemahaman linguistik, pembentukan kata dan kalimat di dalam suatu teks atau dokumen tidak selalu memiliki kesamaan dalam tata kerja dan

pengolahannya. Struktur teks dan kalimat yang ada biasanya terbentuk atas gabungan beberapa kata lainnya dan membentuk suatu kata baru dengan makna baru. Hal inilah yang mendasari metode morfologi Lexeme Based dalam pengetahuan linguistik akan penggunaan bahasa.

Dalam Lexeme Based morfologi, dipercaya bahwa teks dan kalimat selalu terdiri atas beberapa unsur pembentuk lain seperti morfem, frasa, imbuhan dan sebagainya (Purnanto, 2009).

Lexeme Based morfologi terbentuk dengan adanya berbagai jenis imbuhan yang digunakan dalam setiap bahasa yang ada. Dalam Lexeme Based morfologi terdapat aturan yang dapat dirancang untuk menentukan penggunaan teks dan kalimat dengan imbuhannya. Aturan di dalam Lexeme Based dirancang secara individual dan diaplikasikan sesuai dengan kebutuhan dari karakter bahasa yang ada.

Beberapa aturan yang biasa digunakan seperti :
- *Suffix-affix removal* : memiliki kegunaan untuk menghapus imbuhan
- *Suffix replenishment* : memiliki kegunaan untuk melengkapi kebutuhan karakter kata
- *Refining* : memiliki kegunaan untuk memperbaiki struktur kata dan ejaan

Metode yang ada dalam Lexeme Based melakukan pendekatan *item - process* (Bloomfield, 1933) yang berarti bahwa setiap kata dan teks yang ada, disebut sebagai item, merupakan hasil dari pengaplikasian beberapa aturan pemberian imbuhan dan pemrosesan kalimat, sehingga untuk melakukan proses *stemming* terhadap kata-kata yang ada, terlebih dahulu harus ditentukan aturan pemberian imbuhan dan pembentukan pengetahuan linguistik akan jenis teks dan kalimat tersebut.

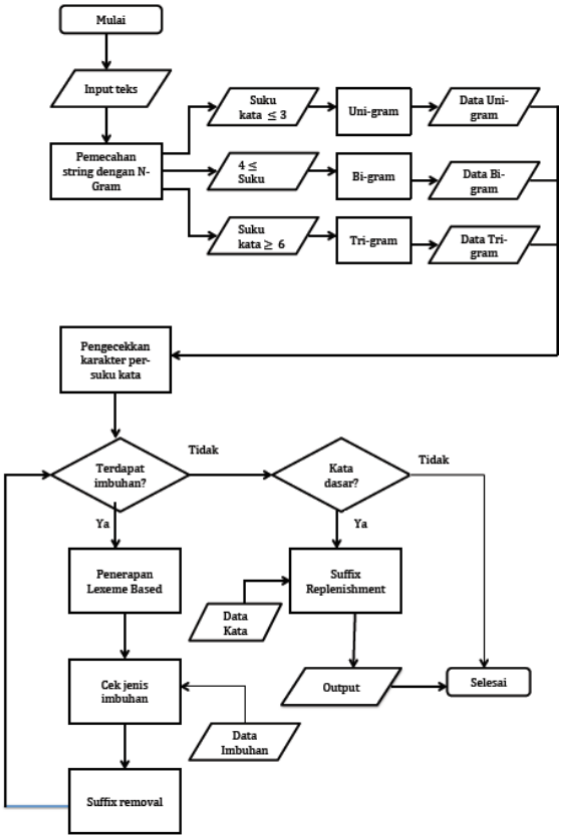
3. ANALISIS KEBUTUHAN DAN PERANCANGAN SISTEM

3.1 Deskripsi dan Analisis Sistem

Bahasa Korea merupakan bahasa aglutinatif yang memiliki banyak bentuk morfem imbuhan untuk mendeskripsikan jenis tata bahasa yang berbeda dengan makna yang beragam. Dengan adanya keragaman ini, dibutuhkan suatu metode khusus yang dapat membantu untuk mempelajari bahasa Korea dalam menentukan makna kata yang sebenarnya dan untuk mempelajari penggunaan tata bahasa dari kata dasar tersebut. Pada penelitian kali ini, teknik yang digunakan adalah penggabungan dari metode N-gram dan Lexeme Based. Melalui metode N-gram kata kerja berimbuhan dipecah menjadi beberapa bentuk potongan karakter string dan setelah itu diterapkan metode Lexeme Based dengan beberapa aturan tertentu sehingga bisa didapatkan hasil akhir berupa kata dasar dan makna dari kata tersebut, serta jenis tata bahasa dan imbuhan apa yang digunakannya.

Sistem yang dibangun pada penelitian ini merupakan sebuah sistem yang akan melakukan proses stemming atau pemotongan kata dari inputan kata berbahasa Korea, sehingga nantinya akan mengeluarkan output berupa kata dasar yang dilengkapi dengan makna dari kata tersebut dan penjelas mengenai imbuhan yang melekat pada kata itu.

Rancangan proses kerja yang terdapat pada sistem dapat dilihat pada diagram berikut :



Gambar 1 Diagram Rancangan Sistem

Sebelum proses stemming dilakukan, kata inputan akan melewati proses pengecekan jumlah string atau karakter yang terdapat di dalam kata tersebut, sehingga dapat ditentukan nantinya jenis metode N-Gram seperti apa yang dapat diimplementasikan (Uni-gram, Bi-gram, atau Tri-gram). Pada tahap awal, setelah proses pengecekkan jumlah karakter, kata berimbuhan tersebut akan dipisahkan menggunakan metode N-gram. Jika terdapat kurang dari 3 (tiga) karakter di dalam kata tersebut maka akan diproses dengan Uni-gram, kemudian jika terdapat lebih dari 3 (tiga) karakter dan kurang dari atau sama dengan 6 (enam) karakter maka akan diproses dengan Bi-gram, lalu jika terdapat lebih dari 6 (enam) karakter, maka akan diproses dengan Tri-gram.

Setelah proses dengan metode N-Gram ini selesai, maka akan dilanjutkan proses stemming menggunakan metode Lexeme Based. Pada tahap ini hasil pemrosesan kata pada tahap sebelumnya akan diproses kembali dengan pengimplementasian rule yang ada pada metode Lexeme Based, yaitu Suffix Removal dan Suffix Replenishment. Dari hasil pemisahan karakter menggunakan N-Gram, karakter yang ada di dalam kata akan dicek satu persatu dan dicocokkan dengan karakter imbuhan yang ada pada daftar imbuhan, lalu saat ditemukan kecocokkan maka akan diproses dengan Suffix Removal, sehingga karakter imbuhan dipisahkan sepenuhnya dari kata dasar. Kemudian dilanjutkan dengan proses pengecekkan kata dasar menggunakan rule Suffix Replenishment. Pada tahap ini, hasil pemrosesan dari tahap Suffix Removal akan kembali dicocokkan pada database untuk melihat

apakah sudah terbentuk kata dasar yang benar, jika belum akan kembali diulangi proses Suffix Removal hingga karakter imbuhan benar-benar sudah terpisah dari kata dasar tersebut. Dalam tahap Suffix Replenishment, jika kata tersebut sudah benar dan cocok termasuk ke dalam bentuk kata dasar maka akan dilanjutkan dengan proses pencarian makna dari kata dasar tersebut dan dilakukan proses pencarian makna dari karakter imbuhan yang sebelumnya sudah dipisahkan, sehingga outputnya akan menjadi suatu pembelajaran baru bagi pengguna, yaitu menampilkan bentuk kata dasar dari kata kerja berimbuhan dan disertai dengan makna dari kata tersebut, kemudian juga dilengkapi dengan informasi penjelas mengenai imbuhan yang melekat sebelumnya pada kata tersebut.

Input

Kata : 먹을수있습니다 [môgeulsuisseubnida]

Proses 1

Pengecekkan jumlah suku kata :

먹을수있습니다 [môgeulsuisseubnida]
➔ 먹 [môg] + 을 [eul] + 수 [su] + 있 [iss] + 습 [seub] + 니 [ni] +다 [da] = 7 suku kata

Proses 2

Menentukan proses stemming dengan N-Gram :
Suku kata ≥ 7 ➔ Tri-Gram

먹을수있습니다 [môgeulsuisseubnida]
➔ 먹을수[môgeulsu] , 을수있[eulsuiss] , 수있습 [suissseub] , 있습니 [isseubni] , 습니다 [seubnida]

Proses 3

Melakukan proses pengecekkan karakter imbuhan :
Suku kata ≥ 7 ➔ Tri-Gram

먹을수있습니다 [môgeulsuisseubnida]
➔ 먹을수[môgeulsu] , 을수있[eulsuiss] , 수있습 [suissseub] , 있습니 [isseubni] , 습니다 [seubnida]

Proses 4

Melakukan proses pemisahan karakter imbuhan :

먹을수있습니다 [môgeulsuisseubnida]
➔먹 [môg] – imbuhan : 을수있습니다 [eulsuisseubnida]

Proses 5

Melakukan proses pengecekkan kata dasar dan suffix replenishment :

먹을수있습니다 [môgeulsuisseubnida] ➔먹다 [môgda]
Imbuhan : 을수있습니다 [eulsuisseubnida]

Proses 6

Melakukan proses pencarian arti kata dasar :

먹을수있습니다 [môgeulsuisseubnida] ➔먹다 [môgda] : makan
Imbuhan : 을수있습니다 [eulsuisseubnida]

Proses 7

Melakukan proses pencarian arti imbuhan dan jenisnya:

먹을수있습니다 [môgeulsuisseubnida] ➔먹다 [môgda] : makan

Imbuhan : 을수있습니다 [eulsuisseubnida] ➔ dapat melakukan sesuatu (formal)
Jenis imbuhan : Ability expression in honorific

Gambar 2 Alur Proses Sistem

4. PENGUJIAN DAN ANALISIS

4.1 Analisis Pengaruh Jumlah Suku Kata terhadap Sistem

Pada pengujian ini, dibentuk kelompok data uji dengan karakteristik data yang berbeda, yaitu kata kerja berimbuhan dengan suku kata berjumlah 1 (satu) hingga 3 (tiga) sebagai Tipe 1, kemudian kata kerja berimbuhan yang memiliki suku kata berjumlah 4 (empat) hingga 6 (enam) sebagai Tipe 2, dan kata kerja berimbuhan yang memiliki suku kata berjumlah lebih dari 6 (enam) sebagai Tipe 3. Pengujian ini bertujuan untuk melihat pengaruh jumlah suku kata pada kata berimbuhan tersebut terhadap hasil dari seluruh proses pengolahan data. Berikut adalah hasil pengujian sistem yang direpresentasikan oleh persentase nilai akurasi :

Tabel 1 Nilai Akurasi Hasil Uji Skenario 1

DATA UJI	AKURASI			
	P1	P2	P3	Rata-rata
Tipe 1	91%	91%	91%	91%
Tipe 2	89%	89%	89%	89%
Tipe 3	83%	83%	83%	83%

Gambar 3 Grafik Nilai Akurasi Hasil Uji

Pada percobaan pertama untuk Tipe 1, didapatkan akurasi yang cukup tinggi, hal ini dikarenakan oleh jumlah suku kata yang masih sedikit dan

pemrosesan Uni-gram yang mempermudah proses pada metode *Lexeme* Based sehingga dapat dihasilkan data yang sesuai.

Pada pengujian dengan Tipe 2, didapatkan nilai akurasi yang lebih rendah dibandingkan dengan Tipe 1, hal ini dikarenakan oleh jumlah suku kata yang semakin banyak membuat kerja sistem lebih sulit dan kompleks serta penggunaan metode Bi-gram pada kata berimbuhan yang membuat pembentukan kata dasar dan imbuhan lebih rumit dibandingkan metode Uni-gram pada Tipe 1. Dalam kata kerja berimbuhan berbahasa Korea, semakin banyak jumlah suku kata maka semakin banyak kemungkinan pembentukan imbuhan dan kata dasar yang bisa terjadi, sehingga nilai akurasi dalam pengujian Tipe 2 mengalami penurunan dibandingkan dengan Tipe 1.

Pada pengujian Tipe 3, didapatkan hasil akurasi yang ternyata lebih rendah dibandingkan dengan pengujian oleh Tipe 1 dan Tipe 2. Pada pengujian ini, disediakan kata berimbuhan yang memiliki suku kata berjumlah lebih dari 6 (enam), hal ini membuat proses pengolahan di dalam sistem menjadi lebih panjang dan kompleks, dikarenakan semakin banyaknya suku kata atau karakter yang perlu diproses oleh sistem untuk pengelompokkan kata dasar dan imbuhan, sehingga nilai akurasi menjadi lebih rendah.

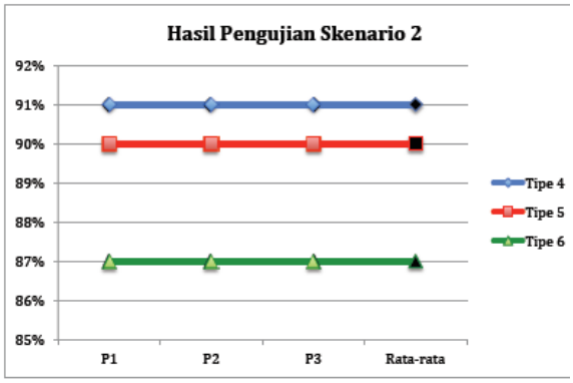
Dari keseluruhan pengujian pada Skenario 1 dapat diketahui bahwa semakin banyak jumlah suku kata atau karakter di dalam satu kata dapat membuat proses pengolahan data menjadi semakin kompleks dan sulit. Kemudian perbedaan nilai akurasi yang terdapat untuk masing-masing percobaan terjadi karena adanya beberapa perbedaan khusus dalam karakteristik huruf dan suku kata dari kata berimbuhan yang ada, sehingga sistem masih belum bisa memproses data secara maksimal untuk keseluruhan jenis kata berimbuhan namun sistem sudah dapat dinilai sebagai sistem yang baik karena secara keseluruhan pengolahan data dalam sistem dapat menghasilkan data yang diinginkan.

4.2 Analisis Pengaruh Karakter Asing terhadap Sistem

Pengujian pada Skenario 2 ini dilakukan dengan tujuan untuk memeriksa pembentukan imbuhan dan kata dasar dari hasil pemecahan suku kata di dalam kata berimbuhan tersebut. Pengujian ini memiliki 3 (tiga) kelompok data yang mirip dengan data pada Skenario 1, namun seluruh data pada pengujian ini telah diubah beberapa suku kata dan karakternya secara acak, sehingga sistem akan menerima karakter asing yang berbeda dari yang seharusnya diproses, pengujian ini juga dilakukan sebanyak 3 (tiga) kali percobaan. Berikut adalah hasil pengujian sistem yang direpresentasikan oleh persentase nilai akurasi :

Tabel 2 Nilai Akurasi Hasil Uji Skenario 2

DATA UJI	AKURASI			
	P1	P2	P3	Rata-rata
Tipe 4	91%	91%	91%	90%
Tipe 5	90%	90%	90%	90%
Tipe 6	87%	87%	87%	87%



Gambar 4 Grafik Nilai Akurasi Hasil Uji

Dalam Tipe 4, disisipkan 1 (satu) suku kata atau karakter asing secara acak di dalam kata berimbuhan untuk mencari tau apakah sistem dapat menjalankan proses pembentukan kata dasar dan imbuhan dengan benar. Dalam pengujian ini disisipkan karakter asing untuk memeriksa apakah proses pada metode Lexeme Based dapat berjalan, dan ternyata sistem menolak untuk memproses kata berimbuhan tersebut yang berarti pengolahan data di dalam sistem berjalan dengan benar pada Tipe 4.

Pengujian selanjutnya dilakukan dengan Tipe 5, yaitu memiliki 1 (satu) jenis suku kata atau karakter asing didalamnya. Dengan hasil akurasi yang didapatkan pada pengujian Tipe 5 ini bisa disimpulkan bahwa sistem masih bisa memproses data dengan baik meskipun jumlah suku katanya lebih kompleks dengan sisipan satu karakter asing.

Pengujian terakhir pada Skenario 2 yaitu dilakukan dengan data Tipe 6, yang terdiri atas suku kata yang berjumlah lebih dari 6 (enam) dan memiliki 2 (dua) suku kata atau karakter asing sebagai sisipan. Disini hasil akurasi pengujian menjadi yang paling rendah dibandingkan dengan pengujian pada Tipe 4 dan Tipe 5, dikarenakan jumlah suku kata yang semakin banyak dengan karakter sisipan yang juga bertambah. Proses pengolahan menjadi semakin rumit sehingga kesalahan dalam pengolahan data pun terjadi lebih banyak pada pengujian dengan Tipe 6.

Dari keseluruhan pengujian dengan Skenario 2 ini dapat disimpulkan bahwa dengan menyisipkan karakter asing didalam kata berimbuhan dapat memastikan bahwa proses Lexeme Based dalam sistem dapat menjalankan prosesnya dengan sesuai, karena saat terdapat karakter asing pada kata berimbuhan, sistem secara otomatis akan memeriksa data corpus secara menyeluruh satu persatu untuk melihat kecocokan yang mungkin ada antara data yang diproses dengan corpus hingga pada akhirnya sistem dapat mengenali dan mengklasifikasikan bentuk kata dasar dan imbuhan yang seharusnya dengan baik saat ditemukannya karakter asing yang tidak sesuai dengan data.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang didapatkan dari penelitian yang telah dilakukan adalah sebagai berikut :

1. Metode N-Gram dan Lexeme Based dapat menjadi teknik gabungan yang sesuai untuk melakukan pencarian kata dasar pada teks berbahasa Korea
2. Sebagai bahasa aglutinatif, kata dalam bahasa Korea memiliki banyak jenis pelekatan imbuhan dengan beragam jenis karakter dan arti yang berbeda.

5.2 Saran

Saran yang dapat diberikan dari penelitian yang telah dilakukan adalah sebagai

berikut :

1. Proses pencarian kata bisa diperluas untuk jenis kata lainnya dalam bahasa Korea selain kata kerja.
2. Proses input data bisa dilakukan secara kolektif untuk jenis kata berbeda
3. Menambah jenis imbuhan yang bisa diproses dalam sistem

DAFTAR PUSTAKA

[1] Bloomfield, Leonard. 1933. *Language*. New York : Henry Holt.

[2] Chaer, Abdul. 2007. *Linguistik Umum.* Jakarta : Rineka Cipta.

[3] Cho, Sehyeong, Seung Soo Han. 2009. *Automatic Stemming for Indexing of an Agglutinative Language.* Gyeonggi : Myongji University.

[4] Choi, Jeon Seung, *et al*. 2009. *Gugohakeui Ihae.* Gyeonggido : Thaehaksa

[5] Dickinson, Markus, Ross Israel dan Sun Hee Lee. 2010. *Building a Korean Web Corpus for Analyzing Learner Language.* Indiana : Indiana University.

[6] Kim, Minjung & Kwon Hyukchul. 1996. *Rule-based Approach to Korean Morphological Disambiguation Supported by Statistical Method.* Busan : Busan National University.

[7] Lee, Changyeol. 1999. *Local Grammar-based Lexical Stemmer for Korean Language.* Venezia : KERIS.

[8] Lee, Ikseop. 2006. *Hangugo Munbeob (A Korean Grammar).* Seoul : Seoul National University Press.

[9] Lee, Ju Haeng. 2004. *Hangugo Munbeobeui Ihae.* Seoul : Doseochulpan Worin.

[10] Lee, Junho & Ahn Jeongsu. 1996. *Using n-Grams for Korean Text Retrieval.* Daejeon : Korea Institute of Science and Technology.

[11] Lee, Kwan Kyu. 2007. *Hakkyo Munbeobron.* Seoul : Doseochulphan Worin.

[12] Purnanto, Dwi. 2009. *Peranan Leksem dan Kata Dalam Studi Morfologi* (Tesis). Surabaya : Sastra Universitas Negeri Semarang.