Peningkatan Algoritma Porter *Stemmer* Bahasa Indonesia berdasarkan Metode Morfologi dengan Mengaplikasikan 2 Tingkat Morfologi dan Aturan Kombinasi Awalan dan Akhiran

Putu Bagus Susastra Wiguna[1], Bimo Sunarfri Hantono[2]

***Abstract*— Stemmer has been used in document processing like: information retrieval, question answering, spell checking, language translator, document clustering, document classification. Stemmer method based on word morphology has some lack such as: incorrect prefix removal on root words beginning with the letter "k", "t", "s" and "p", Incorrect suffix removal especially for "-kan" and "-an" suffix. To handle these problems, this research proposes a stemmer that uses two level morphology to root word beginning with the letter "k", "t", "s", "p" and use prefix and suffix combination rules to remove suffix on a word. Example: "di-" as the prefix should only be paired with "kan-" as the suffix and should not be paired with "-an" as the suffix. The experiments showed that the proposed stemmer accuracy was 95.5%, better than the earlier stemmer based on word morphology. The accuracy of earlier stemmer based on word morphology was 82.5%.**

***Intisari*— *Stemmer* telah digunakan secara luas dalam pengolahan dokumen elektronik seperti: sistem temu kembali informasi (*information retrieval*), *question answering*, pemeriksaan ejaan, mesin penerjemah, *clustering* dokumen, klasifikasi dokumen. Metode *stemmer* dengan menggunakan morfologi suatu kata memiliki beberapa kekurangan seperti tidak tepat menghilangkan awalan pada kata dasar yang berawalan huruf "k", "t", "s" dan "p" serta tidak tepat dalam menghilangkan akhiran terutama untuk akhiran "-kan" dan "-an." Untuk menyelesaikan masalah ini, penelitian ini menawarkan penggunaan 2 tingkat morfologi pada kata dasar berawalan huruf "k", "t", "s" dan "p" serta menggunakan aturan kombinasi awalan dan akhiran untuk menghilangkan akhiran pada suatu kata seperti awalan "di-" hanya boleh dipasangkan dengan akhiran "-kan" dan tidak boleh dengan akhiran "-an" Hasil dari penelitian ini adalah *stemmer* yang memiliki tingkat akurasi 95,5%, lebih baik dibandingkan *stemmer* sebelumnya yang menggunakan algoritma berdasarkan morfologi suatu kata. *Stemmer* sebelumnya yang menggunakan algoritma berdasarkan morfologi suatu kata memiliki tingkat akurasi 82,5%.**

***Kata Kunci*— *Stemmer*, 2 tingkat morfologi, kombinasi awalan dan akhiran**

[1] *Mahasiswa Pascasarjana Teknik Elektro, Jurusan Teknik Elektro dan Teknologi Informasi Fakultas Teknik Universitas Gadjah Mada, Jln. Grafika 2 Yogyakarta 55281 INDONESIA (telp:0274-552305; e-mail: bagussusastra_ti08@mail.ugm.ac.id)*

[2] *Dosen Jurusan Teknik Elektro dan Teknologi Informasi Fakultas Teknik Universitas Gadjah Mada, Jln. Grafika 2 Yogyakarta 55281 INDONESIA (telp: 0274-552305; fax: 0274-552305; e-mail: bhe@ugm.ac.id)*

I. PENDAHULUAN

Stemming telah digunakan secara luas dalam pengolahan dokumen elektronik. *Stemming* digunakan dalam beberapa bidang seperti: sistem temu kembali informasi (*information retrieval*), *question answering* (QA), pemeriksaan ejaan, mesin penerjemah, *clustering* dokumen, klasifikasi dokumen dan lain-lain. *Stemming* [1] adalah prosedur komputasi yang mengubah kata menjadi bentuk asalnya (*stem*) dengan mencari awalan, akhiran dan menghapusnya berdasarkan aturan suatu bahasa. Hasil dari proses *stemming* disebut dengan *token*. Salah satu keuntungan menggunakan *stemming* dalam pengembangan sistem temu kembali informasi (*information retrieval*) adalah: efisiensi dan *index file* yang sudah terkompresi. Misal seperti ini: seorang pencari memasukan *term stemming* sebagai bagian dari *query*. Hal itu menunjukkan bahwa orang tersebut juga tertarik pada *stemmed* dan *stem*. Tanpa proses *stemming*, kata "*stemming*", "*stemmed*" dan "*stem*" adalah sesuatu yang berbeda. Dengan proses *stemming* maka setiap kata yang memiliki akar kata yang sama masih dapat disamakan meskipun tidak memiliki kata-kata yang persis sama.

Morfologi [2] adalah suatu penelitian yang mempelajari tentang cara suatu kata dibangun dari unit-unit yang lebih kecil. Dalam bahasa Inggris kata "*kind*" terdiri dari satu unit terkecil yang biasa disebut dengan kata dasar sedangkan kata "*players*" terdiri dari 3 unit terkecil yaitu: "*play*", "*-er*" dan "*-s*". Unit terkecil "*kind*" dan "*play*" dapat berdiri sendiri sebagai kata sedangkan imbuhan "*-er*" dan "*-s*" harus dilekatkan dengan unit terkecil lainnya agar dapat menjadi sebuah kata.

Dalam bahasa Indonesia ada beberapa penelitian yang dilakukan yang berhubungan dengan *stemmer* bahasa Indonesia seperti [3]. *Stemmer* yang dikerjakan oleh Tala di [3] adalah *stemmer* yang digunakan aplikasi *text mining* bahasa Indonesia. Semua penelitian yang telah dilakukan mengubah bentuk kata menjadi kata dasar dengan berdasarkan bentuk morfologi dari suatu kata. Hal ini tidak tepat jika digunakan untuk mencari kata dasar yang secara struktur yang tidak biasa misal: mengacaukan memiliki struktur kata yg berbeda dengan mengadukan, mengamankan. Aturan awalan dan imbuhan yang dilakukan pada *stemmer* [3] juga mengalami beberapa kesalahan jika kata yang akan dicari bentuk dasarnya memiliki akhiran kata ganti milik seperti "-mu", "-nya", "-ku". Sebagai contoh kata paku menjadi "pa"+"ku". Hal ini dikarenakan akhiran kata ganti milik lebih dulu dieksekusi dan tidak ada aturan pengecualian untuk kata-

kata dasar yang memiliki akhiran kata ganti milik. Kelemahan lainnya pada *stemmer* [3] adalah tidak menyediakan kamus dari daftar kata-kata dasar sehingga menyebabkan kesalahan pada kata seperti: "peranakan" yang diproses menjadi "per"+"ana"+"kan". Hal ini dikarenakan dalam algoritma yang digunakan lebih dulu melihat akhiran –kan daripada akhiran "–an".

Penelitian tentang porter *stemmer* juga dilakukan oleh Baskoro dari Ilmu Komputer Universitas Gadjah Mada. Baskoro melakukan proses *stemming* dengan melihat struktur dari suatu kata [4]. Penelitian yang dilakukan oleh Baskoro lebih dulu melihat akhiran partikel seperti "–kah", "-lah", "-tah" dan kata ganti milik seperti "-ku", "-mu", "-nya" dibandingkan melihat awalan dari suatu kata [4]. Algoritma ini menjadi tidak efisien karena diperlukan masing-masing 2 tabel di *database* untuk kata berakhiran partikel dan kata ganti milik. Selain itu kekurangan algoritma pada [4] adalah algoritma ini tidak memperhitungkan kata dasar yang berawalan huruf "k", "t", "s", "p" dengan benar. Dalam bahasa Indonesia kata yang diawali dengan huruf "k", "t", "s", "p" jika diberi awalan "me-" atau "pe-" maka huruf depannya akan melebur. Sebagai contoh: "mengacau" menjadi "me"+ "kacau". Jika kata "mengacau" dicari kata dasarnya dengan menggunakan porter *stemmer* dari [4] maka kata dasar yang muncul adalah "acau". Kata ini tentu tidak bisa disamakan dengan kata "mengantar" menjadi "me" + "antar" walaupun secara struktur sama.

Penelitian ini mengembangkan suatu *stemmer* yang mencari kata dasar dari suatu kata berdasarkan morfologi suatu kata dengan mengaplikasikan 2 tingkat morfologi dan aturan kombinasi awalan dan akhiran. Penggunaan 2 tingkat morfologi dapat menghindari kesalahan pada saat menghilangkan awalan pertama pada kata dasar yang berawalan huruf "k", "t", "s" dan "p". Aturan kombinasi awalan dan akhiran dapat digunakan untuk menghindari kesalahan pada saat menghilangkan akhiran "-kan" dan "-an" yang tidak dapat dilakukan pada penelitian sebelumnya [3].

II. STRUKTUR KATA DALAM BAHASA INDONESIA

A. Tata Bahasa Indonesia Berdasarkan Strukturnya

Berdasarkan strukturnya, kata dalam bahasa Indonesia dapat dilekati 5 jenis imbuhan yang berbeda yaitu: awalan, sisipan, akhiran, kata ganti milik dan partikel. Tabel I menunjukkan daftar imbuhan yang dapat ditambahkan pada kata dalam bahasa Indonesia.

TABEL I
IMBUHAN PADA KATA DALAM BAHASAINDONESIA [5]

	Morfem	Contoh
Awalan	me-, di-, be-, pe-, ke-, ter-, se-	menabung, dipukul, berapi
Sisipan	-em-, -el-, -er-	gerigi, telunjuk
Akhiran	-kan, -an, -i, -isme,-isasi	makanan,
Kata ganti milik	-ku, -mu, -nya	punyanya, bukuku
Partikel	-lah, -kah, -tah, -pun	masihkah, bacalah

Struktur dari suatu kata dalam bahasa Indonesia dirumuskan [3]:

[prefix1] + [prefix2] + root + [suffix] + [possessive pronoun] + [particle}

Struktur suatu kata bahasa Indonesia pada [3] memperlihatkan bahwa suatu kata dalam bahasa Indonesia dibangun dari suatu kata dasar dengan menggunakan berbagai operasi morfologi meliputi menggabungkan, menambahkan imbuhan dan pengulangan [6]. Bentuk pengulangan suatu kata dapat dibagi menjadi 2 jenis yaitu: pengulangan penuh dan pengulangan sebagian. Contoh dari pengulangan penuh adalah "buku-buku" yang berasal dari kata dasar "buku", "mata-mata" yang berasal dari kata dasar "mata" Pengulangan sebagian meliputi pengulangan dengan menambahkan imbuhan pada kata dasar seperti: "buah-buahan" dengan kata dasar "buah", "bertingkat-tingkat" dengan kata dasar "tingkat". Tidak semua kombinasi awalan dan akhiran dapat digunakan bersama-sama. Ada beberapa kombinasi awalan dan akhiran yang tidak diijinkan dalam tata bahasa Indonesia seperti yang terlihat pada Tabel II

TABEL II
KOMBINASI AWALAN DAN AKHIRAN YANG TIDAK DIIJINKAN [3]

Awalan	Akhiran
Ber	I
Di	An
Ke	i|kan
Meng	An
Peng	i|kan
Ter	An

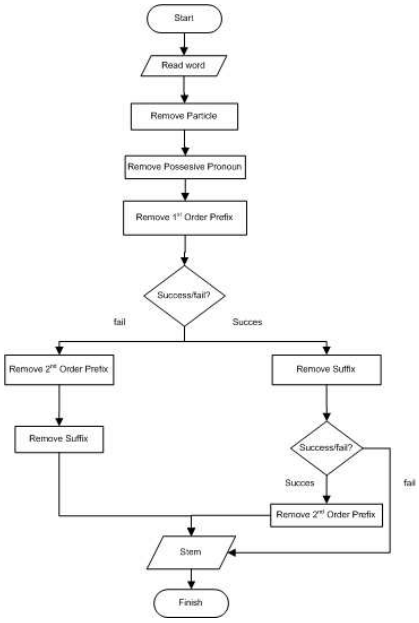
Secara umum morfologi dibagi menjadi 2 proses yaitu *inflectional process* dan *derivational process* [6]. *Inflectional process* adalah proses perubahan suatu kata yang tidak mengubah jenis kata dasarnya. Misal: kata "memukuli" berasal dari kata dasar "pukul". Kata "pukul" merupakan sebuah kata kerja dan kata "memukuli" juga merupakan kata kerja sehingga tidak terjadi perubahan jenis kata pada proses tersebut. *Derivational process* adalah suatu proses perubahan kata dasar yang mengubah jenis kata dasarnya. Misal: kata "pemukul" merupakan kata benda yang memiliki kata dasar "pukul" yang merupakan kata kerja.

B. Dua Tingkat Morfologi

Tidak semua proses pembentukan kata dari kata dasar bisa diselesaikan dengan satu tingkat morfologi. Contoh pembentukan kata dengan penambahan imbuhan pada kata dasar dengan satu tingkat morfologi adalah "mem"+"baca" menjadi "membaca", "men"+"cari" menjadi "mencari". Penambahan imbuhan pada kata dasar untuk membentuk kata baru dengan mengubah fonem dari kata dasar tidak bisa diselesaikan dengan satu tingkat morfologi.Contoh kata yang tidak bisa diselesaikan dengan satu tingkat morfologi adalah kata "memutar" berasal dari kata dasar "putar" yang mendapat imbuhan "men-". Untuk menyelesaikan masalah ini maka diperlukan 2 tingkat morfologi untuk menyelesaikan masalah ini. Penggunaan 2 tingkat morfologi adalah cara lain untuk mendeskripsikan fonem pada *finite-state term* [7].

III. METODE *STEMMING*

Suatu *stemmer* dibangun dengan asumsi bahwa tidak ada suatu kata yang bermakna ganda [3]. Proses yang dilakukan adalah: menghilangkan partikel, menghilangkan kata ganti milik, menghilangkan awalan pertama, menghilangkan awalan kedua, menghilangkan suffix dan selanjutnya ditemukan kata dasar dari suatu kata [3]. Implementasi porter *stemmer* yang dilakukan dapat dilihat pada Gbr. 1 [3].



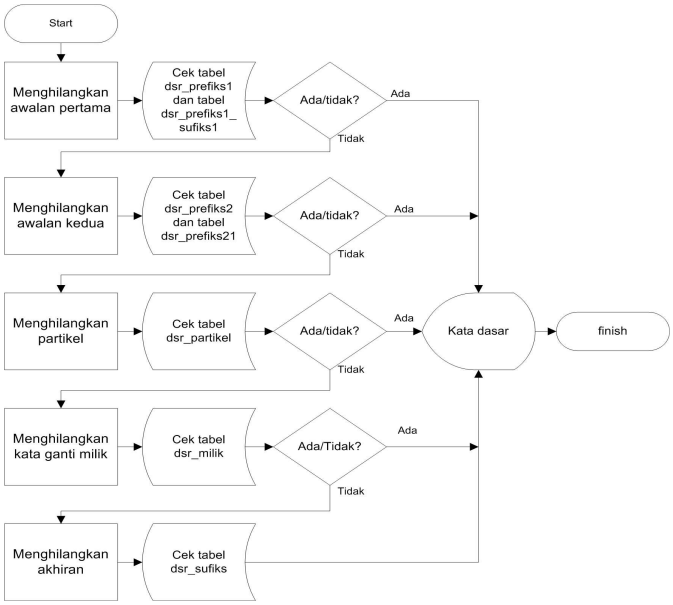
Gbr. 1 Algoritma yang digunakan pada [3] dan [4]

Seperti yang ditunjukan pada Gbr. 1 dapat dilihat ada beberapa kekurangan pada algoritma tersebut. Kekurangan pada algoritma yang ditunjukkan pada Gbr. 1 yang akan ditingkatkan pada penelitian ini adalah:

- Algoritma pada Gbr. 1 lebih dulu menghilangkan partikel dan kata ganti milik dibandingkan awalan 1 dan awalan 2. Algoritma ini menggunakan *database* untuk menyimpan kata yang tidak perlu diproses karena merupakan kata-kata pengecualian tiap prosesnya. Misal pada proses menghilangkan partikel. Proses ini menghilangkan partikel “-kah”, “-pun”, “-lah”, “-pun”. Pada proses ini ada beberapa kata yang tidak perlu diproses karena kata tersebut merupakan kata dasar yang mengandung partikel tersebut seperti kata “nikah”. Fonem atau partikel “-kah” pada kata “nikah” tidak perlu dihilangkan karena merupakan bagian pembentuk suatu kata dasar dan kata “nikah” dimasukkan ke dalam *database*. Diperlukan 2 tabel (tabel yang berisi kata dasar berpartikel dan tabel yang berisi kata dasar berpartikel yang mendapatkan awalan), jika lebih dahulu menghilangkan partikel dibandingkan menghilangkan awalan. Misal kata “menikah” yang berasal dari kata “nikah”. Partikel –kah pada kata “menikah” akan dihilangkan jika hanya menggunakan 1 tabel (tabel kata dasar berpartikel) sehingga akan menghasilkan “meni” untuk dilanjutkan pada proses selanjutnya. Hal ini dapat diselesaikan dengan menghilangkan awalan terlebih dahulu daripada harus menambah 1 tabel yang tentunya akan tidak efisien.
- Algoritma pada Gbr. 1 tidak memperhitungkan 2 tingkat morfologi yang berlaku pada kata dasar yang berawalan huruf “k”, “t”, “s”, “p” dengan tepat jika diberikan awalan “pe-” dan “me-”. Kata “menikah” memiliki morfologi yang sama dengan “menulis” akan tetapi memiki huruf awal yang berbeda pada kata dasarnya. Kata “menikah” berasal dari kata dasar “nikah” dan berawalan huruf “n”. Sedangkan kata “menulis” berasal dari kata dasar “tulis” yang berawalan huruf “t”. Proses pada kata “menikah” tidak bisa diterapkan pada kata “menulis” karena bisa menghasilkan kata dasar yang salah. Oleh karena itu diperlukan 2 tingkat morfologi yang tepat untuk menyelesaikan masalah ini.
- Penerapan aturan kombinasi awalan dan akhiran belum berjalan dengan baik, terbukti dengan adanya kesalahan dalam menghilangkan akhiran pada suatu kata terutama akhiran “-kan” dan “-an”.

Untuk menunjang proses *stemming* dapat dilakukan dengan baik maka diperlukan *database* kata yang terdiri dari 7 tabel yang menjadi kamus kata-kata pengecualian untuk tiap prosesnya. 7 tabel yang digunakan adalah tabel dsr_milik, tabel dsr_partikel, tabel dsr_prefiks1, tabel dsr_prefiks1_sufiks1, tabel dsr_prefiks2, tabel dsr_prefiks21, tabel dsr_sufiks.

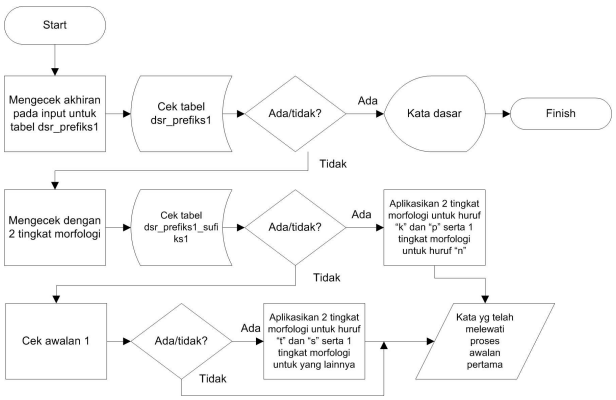
Secara umum proses *stemming* dibagi menjadi 5 bagian yaitu: menghilangkan awalan pertama (“meng-”, “peng-”, “mem-”, “pem-”, “meny-”, “peny-”, “men-”, “pen-” dan lain-lain), menghilangkan awalan kedua (“ber-”, “per-”, “ter-”, “se-”, “pel-”, dan lain-lain), menghilangkan partikel (“-kah”, “-lah”, “-tah”, “-tah”), menghilangkan kata ganti milik (“-ku”, “-mu”, “-nya”), menghilangkan akhiran (“-kan”, “-an”, “-i”, “-isme”, “-isasi”, “-onal”). Proses *stemming* secara umum yang dilakukan pada penelitian ini dapat dilihat pada **Error! Reference source not found.**.



Gbr. 2 *Flowchart* proses *stemming* secara umum

Pada **Error! Reference source not found.** dapat dilihat tahap pertama adalah menghilangkan awalan pertama. Pada tahap ini perbaikan algoritma dilakukan dengan memperhitungkan 2 tahap morfologi untuk kata dasar yang diawali huruf “k”, “t”, “s” dan “p”.

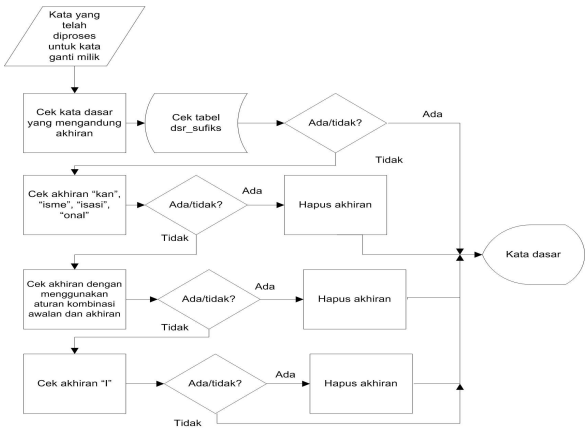
Gbr. 3 menunjukkan 2 tingkat morfologi diterapkan dengan baik pada tahap menghilangkan awalan pertama. Ada 2 tabel yang digunakan pada penelitian ini yaitu tabel dsr_prefiks1 dan tabel dsr_prefiks1_sufiks1. Tabel dsr_prefiks1 digunakan untuk menyimpan kata dasar yang memiliki fonem awalan pertama. Tabel ini berguna agar fonem awalan pertama pada kata tersebut tidak dihilangkan karena merupakan bagian dari kata dasar.Tabel kedua adalah tabel dsr_prefiks1_sufiks1. Tabel ini digunakan untuk menyimpan kata-kata yang harus diproses 2 tingkat morfologi untuk kata dasar yg berawalan huruf “k” dan “p” serta 1 tingkat morfologi untuk kata dasar berawalan huruf “n” Jumlah kata yang dapat diproses 1 tingkat morfologi untuk kata dasar yang berawalan huruf “n”lebih sedikit dibandingkan dengan jumlah kata yang diproses dengan 2 tingkat morfologi untuk kata dasar yang berawalan huruf “t” dan “s”. Oleh karena itu, kata yang disimpan dalam tabel untuk kata dasar berawalan “n”. Contoh kata dasar berawalan huruf “n” yang diproses dengan 1 tingkat morfologi adalah: “menikah”, “menaikkan”, “menyatakan”. Tahap selanjutnya adalah mengecek awalan pertama secara umum. Pada tahap ini juga dilakukan menghilangkan awalan pertama dengan 2 tahap morfologi untuk kata dasar yang berawalan huruf “t” dan “s”.



Gbr. 3 *Flowchart* untuk menghilangkan awalan pertama

Peningkatan juga dilakukan pada tahap menghilangkan akhiran. Peningkatan yang dilakukan pada tahap menghilangkan akhiran adalah menambahkan kemampuan untuk menghilangkan akhiran “-isme”, “-isasi”, “onal”. Selain itu peningkatan juga dilakukan dengan menghilangkan akhiran sesuai dengan aturan kombinasi awalan dan akhiran. Aturan kombinasi awalan dan akhiran yaitu dengan menentukan akhiran yang melekat pada suatu kata dengan melihat awalan yang melekat pada kata tersebut. Dengan menggunakan aturan ini maka kesalahan dalam menghilangkan akhiran dapat diminimalkan terutama pada akhiran “–an” dan akhiran “-kan”.

Gbr. 4 menunjukkan bahwa tahap menghilangkan akhiran menggunakan 1 tabel yaitu tabel dsr_sufiks. Tabel dsr_sufiks adalah tabel yang digunakan untuk menyimpan kata dasar yang mengandung akhiran sebagai fonem terakhir pembentuk kata dasar. Contoh kata dasar yang mengandung fonem akhiran adalah “tangan”, “pantai”. Tangan mengandung fonem akhiran “-an” sedangkan “pantai” mengandung fonem akhiran “-i” sehingga akhiran pada kata tersebut tidak dapat dihilangkan karena sebagai pembentuk kata dasar.



Gbr. 4 *Flowchart* untuk menghilangkan akhiran

Langkah ketiga untuk menghilangkan akhiran diimplementasikan dengan melakukan pengecekan akhiran menggunakan aturan kombinasi awalan dan akhiran seperti yang disebutkan pada Tabel II.

IV. HASIL DAN PEMBAHASAN

Eksperimen dilakukan dengan menggunakan 10 artikel berita dari Universitas Gadjah Mada. 10 Artikel ini telah melewati proses menghilangkan *stop word* sebelum dilakukan proses *stemming*. Dalam suatu artikel atau dokumen terdapat kata yang tidak memiliki informasi dalam jumlah yang besar yang disebut juga dengan *stop word* [8]. Dengan menghilangkan *stop words* maka komputasi dapat menjadi lebih sederhana dan kata yang diproses adalah kata yang benar-benar memiliki nilai informasi. Sebanyak 555 kata yang berbeda baik kata yang berimbuhan maupun kata dasar dari 10 artikel berita Universitas Gadjah Mada. Keseluruhan kata tersebut dicari kata dasarnya dengan menggunakan *stemmer* ini dan dengan menggunakan *stemmer* dari [4] yang menggunakan algoritma dari [3]

Didapat hasil yang berbeda-beda dari 2 *stemmer*. *Stemmer* yang dihasilkan melalui penelitian ini dapat dengan baik mengaplikasikan 2 tingkat morfologi seperti misal pada kata “mengatakan” menjadi “kata”, kata “pemegang” menjadi “pegang”, kata “pengendali” menjadi kata “kendali”. Sedangkan *stemmer* pada [4] tidak dapat melakukan 2 tingkat morfologi dengan baik untuk kata-kata tersebut kata “mengatakan” menjadi “ata”, kata “pemegang” menjadi “gang” dan kata “pengendali” menjadi “endal”. Kesalahan lainnya yang terjadi pada *stemmer* pada [4] adalah kata yang seharusnya tidak diproses 2 tingkat morfologi, diproses dengan 2 tingkat morfologi seperti kata “pemoderasian” menjadi kata “poderasi”. Tidak ada kata dasar “poderasi” dalam Kamus Besar Bahasa Indonesia (KBBI). Di sisi lain *stemmer* pada [4] benar dalam melakukan *stemming* pada

kata “pemerintah” menjadi “perintah” yang menggunakan 2 tingkat morfologi.

Selanjutnya dengan memasukkan kata yang memiliki akhiran seperti “-kan”, “-an”, “-i”. *Stemmer* yang dihasilkan melalui penetilian ini dapat dengan baik menerapkan aturan kombinasi awalan dan akhiran seperti misal kata “kenaikan” menjadi kata “naik”, “tindakan” menjadi “tindak”. Sedangkan pada *stemmer* yang dihasilkan di [4] tidak dapat menerapkan aturan awalan dan akhiran dengan baik terutama jika sistem seharusnya menghilangkan akhiran “-an” bukan “-kan” seperti pada kata “kenaikan” yang diubah menjadi “nai”, kata “tindakan” yang diubah menjadi “tinda”. Selain itu *stemmer* pada [4] tidak memiliki kamus kata yang memadai untuk proses menghilangkan akhiran ini. Sebagai contoh: *stemmer* pada [4] salah menghilangkan fonem “-an” yang membentuk kata “tangan” sehingga kata “tangan” menjadi “tang” jika diproses oleh *stemmer*.

Percobaan selanjutnya yang dilakukan adalah dengan memasukkan kata berulang. *Stemmer* yang dihasilkan melalui penelitian ini belum bisa mencari kata dasar yang berulang seperti: “negara-negara”, “pusat-pusat”, “sebenar-benarnya”. Sedangkan *stemmer* pada [4] bisa dengan baik mendapatkan kata dasar dari kata berulang apabila kata kedua sama dengan kata yang pertama seperti: “negara-negara” menjadi negara, “pusat-pusat” menjadi “pusat”. Namun jika kata pertama tidak sama dengan kata kedua maka *stemmer* pada [4] salah dalam menentukan kata dasar seperti pada kata “sebenar-benarnya” menjadi “sebenarnya”. Hal ini dikarenakan *stemmer* pada [4] hanya memproses kata yang didepan tanda strip (-) atau kata yang pertama.

Stemmer yang dihasilkan melalui penelitian ini dan *stemmer* pada [4] belum bisa mengidentifikasi suatu kata dilekati dengan 2 buah awalan pertama yang berurutan atau 2 buah awalan kedua yang berurutan. Sebagai contoh: kata ”keberhasilan” menjadi kata “berhasil”, kata “berkelanjutan” menjadi kata “kelanjut”.

Stemmer pada [4] belum memiliki *database* yang memadai sebagai kamus kata untuk setiap prosesnya. Sebagai contoh untuk proses menghilangkan awalan pertama seperti kata “disertasi” menjadi “sertasi” dan terutama untuk kata-kata yang memerlukan 2 tingkat morfologi seperti “memecahkan” menjadi “cah”, “pengembangan” menjadi “embang”. Begitu juga untuk proses-proses lainnya seperti kata dasar berpartikel yaitu: “jumlah” menjadi “jum”

Stemmer yang dihasilkan oleh Purwarianti pada [5] belum bisa melakukan *stemming* untuk kata-kata ambigu yang dihasilkan dari kombinasi imbuhan seperti “berupa” (“be”-“rupa” atau “ber”-“upa”), “beragam” (“be”-“ragam” atau “ber”-“agam”). Kata-kata ambigu tersebut telah dapat dicari kata dasarnya dengan benar dengan menggunakan *stemmer* yang dihasilkan melalui penelitian. *Stemmer* yang dihasilkan melalui penelitian ini memiliki *database* yang digunakan untuk membedakan kata-kata ambigu untuk kata dasar yang berawalan “ber- “, “ter-“ dan “per-“

Stemmer bahasa Indonesia yang dihasilkan pada penelitian ini memiliki tingkat akurasi yang lebih baik dibandingkan dengan *stemmer* pada [4]. *Stemmer* pada penelitian ini salah 25 kata sehingga memiliki akurasi mengacu pada (1).

$$\frac{(555-25)}{555} \times 100\% = 95.5\% \quad (1)$$

Persamaan (1) menunjukkan bahwa tingkat akurasi *stemmer* yang dihasilkan pada penelitian ini adalah 95.5%. Sedangkan akurasi untuk *stemmer* pada [4] mengacu pada (2).

$$\frac{(555-97)}{555} \times 100\% = 82.5\% \quad (2)$$

Persamaan (2) menunjukkan bahwa *stemmer* pada [4] memiliki tingkat akurasi yang lebih rendah jika dibandingkan dengan *stemmer* yang dihasilkan pada penelitian ini yaitu 82.5%

Stemmer yang dihasilkan pada penelitian ini dapat dengan baik mengaplikasikan 2 tingkat morfologi. Sebanyak 54 kata dari 555 kata yang digunakan pada penelitian ini merupakan kata yang secara morfologi merupakan kata yang memerlukan 2 tingkat morfologi untuk menghilangkan awalan pertama yang melekat pada kata tersebut. Dari 54 kata, *stemmer* yang dihasilkan pada penelitian ini berhasil mengubah 54 kata tersebut ke bentuk kata dasarnya dengan baik sehingga akurasi *stemmer* ini dalam mengaplikasikan 2 tingkat morfologi mengacu pada (3).

$$\frac{(54-0)}{54} \times 100\% = 100\% \quad (3)$$

Persamaan (3) menunjukkan bahwa *stemmer* pada penelitian ini dapat dengan baik mengaplikasikan 2 tingkat morfologi. Sedangkan akurasi untuk *stemmer* pada [4] dalam mengaplikasikan 2 tingkat morfologi dapat dilihat pada (4).

$$\frac{(54-22)}{54} \times 100\% = 59.2\% \quad (4)$$

Persamaan (3) dan (4) menunjukkan bahwa *stemmer* yang dihasilkan pada penelitian ini lebih baik dalam mengaplikasikan 2 tingkat morfologi dibandingkan dengan *stemmer* pada [4].

Stemmer yang dihasilkan pada penelitian ini juga dapat dengan baik mengaplikasikan aturan kombinasi awalan dan akhiran. Sebanyak 119 kata dari 555 kata merupakan kata yang memiliki awalan dan akhiran “-kan” dan “-an”. Dari 119 kata, *stemmer* yang dihasilkan pada penelitian ini salah 4 kata sehingga akurasi *stemmer* ini dalam mengaplikasikan aturan kombinasi awalan dan akhiran mengacu pada (4).

$$\frac{(118-4)}{118} \times 100\% = 96.6\% \quad (5)$$

Persamaan (5) menunjukkan bahwa *stemmer* pada penelitian ini dapat dengan baik mengaplikasikan aturan kombinasi awalan dan akhiran. Sedangkan akurasi untuk *stemmer* pada [4] dalam mengaplikasikan aturan kombinasi awalan dan akhiran dapat dilihat pada (6).

$$\frac{(118-32)}{118} \times 100\% = 72.8\% \quad (6)$$

Persamaan (5) dan (6) menunjukkan bahwa *stemmer* yang dihasilkan pada penelitian ini lebih baik dalam mengaplikasikan aturan kombinasi awalan dan akhiran dibandingkan dengan *stemmer* pada [4].

V. KESIMPULAN DAN SARAN

Dengan mengaplikasikan aturan 2 tingkat morfologi dapat meningkatkan kemampuan *stemmer* untuk mendapatkan kata dasar yang tepat untuk kata dasar yang berawalan huruf "k", "t", "s" dan "p" dengan akurasi 100%. Tingkat akurasi ini lebih tinggi jika dibandingkan dengan *stemmer* sebelumnya yang memiliki tingkat akurasi 59.2% dalam mengaplikasikan 2 tingkat morfologi. Aturan kombinasi awalan dan akhiran juga dapat meningkatkan kemampuan *stemmer* untuk menentukan akhiran yang melekat pada suatu kata terutama untuk akhiran "-kan" dan akhiran "-an" dengan tingkat akurasi 96.6%. Tingkat akurasi ini lebih tinggi jika dibandingkan dengan *stemmer* sebelumnya yang memiliki tingkat akurasi 72.8% dalam mengaplikasikan aturan kombinasi awalan dan akhiran. Secara umum *stemmer* pada penelitian ini menghasilkan akurasi lebih baik dari *stemmer* sebelumnya yaitu 95.5% berbanding 82.5%.

Metode dengan melihat morfologi suatu kata menimbulkan masalah yaitu dengan kata yang berhomonim, homofon, homograf dan polisemi. Selain itu juga diperlukan metode untuk mengidentifikasi kata dasar dengan benar pada kata yang berulang.

REFERENSI

[1] J. B. Lovins, *Development of a stemming algorithm.* MIT Information Processing Group, Electronic Systems Laboratory, 1968.

[2] D. Jurafsky and J. H. Martin, "Knowledge in Speech and Language Processing," in *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Second Edition., Pearson-Prentice Hall, 2000.

[3] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia." Master of Logic Project Institute for Logic, Language and Computation Universiteit van Amsterdam The Netherlands, 2003.

[4] D. O. Baskoro, H. Malik, and M. H. Anshari, "PORTER *STEMMER* INFORMATION RETRIEVAL." Computer Science Gadjah Mada University, 2012.

[5] A. Purwarianti, "A non deterministic Indonesian *stemmer*," in *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on*, 2011, pp. 1–5.

[6] F. Pisceldo, R. Mahendra, R. Manurung, and I. W. Arka, "A two-level morphological analyser for the indonesian language," in *Australasian Language Technology Association Workshop 2008*, 2008, vol. 6, pp. 142–150.

[7] K. Koskenniemi, *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki Department of General Linguistik Hallituskatu 11-13 SF-00100 Helsinki 10 Finland, 1983.

[8] C. Silva and B. Ribeiro, "The importance of stop word removal on recall values in text categorization," in *Neural Networks, 2003. Proceedings of the International Joint Conference on*, 2003, vol. 3, pp. 1661–1666.