ALGORITMA BARU PEMBENTUKAN KATA DASAR PADA PROSES *STEMMING* BAHASA INDONESIA

Yeni Anistyasari[1], Eko Hariadi[2]
Jurusan Teknik Informatika, Universitas Negeri Surabaya[1]
yenian@unesa.ac.id[1], ekohariadi@unesa.ac.id[2]

ABSTRACT

Stemming is the process of converting a word to its root form by removing affixes including prefix, suffix, prefix and suffix, and infix. Stemming is utilized as the initial process of information retrieval. Several stemming algorithms for Indonesian have been developed, including the Nazief-Adriani (NA) and Confix-Stripping (CS) algorithms. Both of these algorithms have been tested in Indonesian stemming process to form basic words by removing prefix, suffix, and prefix. However, NA and CS algorithms have not explored deeply about removing infix. This research therefore proposes an idea for improving existing stemming algorithms by adding an infix stripping process. A web-based stemming application was developed to test the proposed algorithm. The input for algorithm testing is news from online portal. The generated root-forms is validated by Bahasa Indonesia experts to determine the effectiveness of the proposed algorithm. Validation results from information technology experts indicate that the application is eligible to be utilized. While the validation results of Bahasa Indonesia experts prove that infix stripping algorithm can generate appropriate root forms and in accordance with the rules of Indonesian.

Keywords: *Indonesian Language, Infix, Stemming*

ABSTRAK

Salah satu faktor penting dalam pembentukan kamus elektronik adalah pembentukan kata dasar melalui proses *stemming*. *Stemming* adalah proses pengubahan sebuah kata ke bentuk dasarnya dengan menghilangkan imbuhan yang terdiri dari awalan, akhiran, awalan dan akhiran, dan sisipan. Beberapa algoritma *stemming* untuk Bahasa Indonesia telah dikembangkan, diantaranya adalah algoritma Nazief-Adriani (NA) dan Confix-Stripping (CS). Kedua algoritma ini telah teruji dalam proses *stemming* Bahasa Indonesia untuk membentuk kata dasar dengan menghilangkan awalan, akhiran, dan awalan-akhiran. Namun, algoritma NA dan CS tidak banyak mengeksplorasi proses pelesapan sisipan dalam pembentukan kata dasar. Penelitian ini mengajukan ide untuk memperbaiki algoritma *stemming* yang telah ada dengan menambahkan proses pelesapan sisipan. Untuk menguji algoritma yang diajukan, sebuah aplikasi *stemming* berbasis web dikembangkan. Masukan yang digunakan untuk pengujian algoritma pelesapan sisipan adalah berita dari portal daring. Kata dasar yang dihasilkan aplikasi dari proses stemming divalidasi oleh pakar Bahasa Indonesia untuk menentukan efektifitas algoritma yang diajukan. Hasil validasi dari pakar teknologi informasi menunjukkan bahwa aplikasi layak digunakan untuk proses uji coba. Sedangkan hasil validasi pakar Bahasa Indonesia membuktikan bahwa algoritma pelesapan sisipan dapat menghasilkan kata dasar yang baik dan sesuai dengan aturan Bahasa Indonesia.

Kata Kunci: Bahasa Indonesia, Sisipan, *Stemming*

PENDAHULUAN

Bahasa Indonesia adalah bahasa utama yang digunakan orang Indonesia untuk berkomunikasi baik secara lisan maupun tulisan. Pemilihan kata dalam pengucapan dan dalam menulis kalimat adalah elemen penting. Dalam hal tata Bahasa Indonesia baku, Kamus Besar Bahasa Indonesia adalah referensi utama dalam penggunaan kata dasar. Selain untuk berkomunikasi lisan, penggunaan bahasa Indonesia yang tepat diperlukan saat menulis dokumen formal, jurnal, laporan, dan sebagainya. Dokumen formal yang baik adalah dokumen yang ditulis dalam Bahasa Indonesia dengan tingkat formalitas yang tinggi, menggunakan kosa kata yang tepat, dan menghindari kesalahan tata bahasa dan ejaan (Widjaja and Hansun 2015).

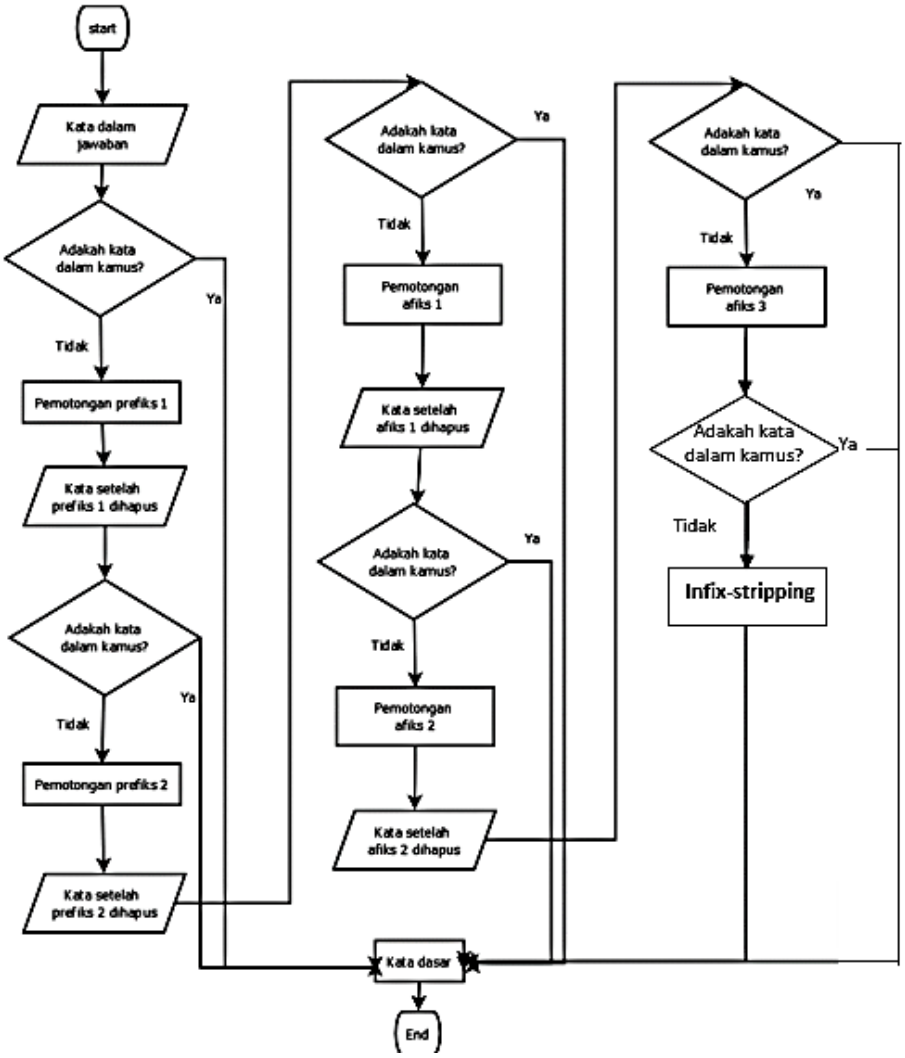
Penelitian tentang aplikasi teknologi informasi untuk Bahasa Indonesia banyak dilakukan, antara lain sistem temu kembali informasi, penarikan kesimpulan dari isi dokumen, kompresi teks, dan klasifikasi teks. Proses penting yang terlibat dalam mengembangkan aplikasi tersebut adalah proses stemming yaitu proses yang dilakukan untuk mengembalikan kata berimbuhan ke bentuk dasarnya dengan cara menghilangkan imbuhan. Imbuhan tersebut berupa awalan, akhiran, awalan dan akhiran, dan sisipan (Winarti, Kerami, and Arief 2017). Dalam klasifikasi teks, stemming bertugas menyederhanakan kata-kata tanpa menghilangkan makna sehingga ukuran kumpulan data akan berkurang. Dalam penarikan kesimpulan dari isi teks, stemming dapat meningkatkan efisiensi. Sedangkan untuk temu kembali informasi, stemming mengurangi jumlah indeks dokumen sehingga meningkatkan kinerja sistem temu kembali informasi, menyediakan varian morfologi dalam mencari istilah, dan meningkatkan akurasi. Sedangkan dalam kompresi teks, stemming memaksimalkan penggunaan kamus dengan menyimpan hanya kata dasar dalam kamus (Widayanto and Huda 2017).

Stemming Bahasa Indonesia pertama dikembangkan oleh Nazief-Adriani kemudian Jelita Asia memperbaiki algoritma tersebut dan menyebut algoritmanya dengan nama confix stripping (CS) stemmer. Ada banyak perbaikan yang dilakukan oleh CS stemmer sehingga menjadi algoritma dengan akurasi stemmer tertinggi (Widayanto and Huda 2017). Pembentukan kata dasar pada kedua algoritma tersebut dilakukan dengan menghilangkan imbuhan. Namun, kedua algoritma tidak membahas secara mendalam tentang sisipan. Seperti diketahui, imbuhan pada Bahasa Indonesia terdiri dari awalan, akhiran, awalan dan akhiran, serta sisipan. Oleh karena itu, penelitian ini mengajukan ide untuk mengembangkan algoritma pelesapan sisipan untuk memperbaikai algoritma stemming yang telah ada. Sumber dokumen masukan untuk proses uji coba algoritma berasal dari portal berita daring. Sebelum dilakukan proses stemming, dokumen yang diunduh diproses tokenisasi untuk mendapatkan tiap kata dari kumpulan kalimat yang panjang.

METODE PENELITIAN

Metode yang digunakan untuk penelitian ini terdiri dari analisa kebutuhan, desain algoritma, implementasi algoritma, dan evaluasi. Tahap analisa kebutuhan yang dilakukan yaitu studi literatur tentang metode stemming Bahasa Indonesia yang telah dikembangkan oleh peneliti lain. Temuan yang diperoleh dari tahap analisa kebutuhan adalah hingga saat ini metode stemming Bahasa Indonesia yang menyertakan pelesapan sisipan belum diteliti secara mendalam. Tahap analisa kebutuhan juga mempelajari aturan sisipan pada Bahasa Indonesia.

Setelah melalui tahap analisa kebutuhan dilakukan desain algoritma. Algoritma yang dikembangkan dinamakan infix-stripping yaitu perbaikan dari algoritma Nazief dan Adriani dan algoritma CS, seperti ditunjukkan Gambar 1.



Gambar 1. Algoritma Infix-stripping

Algoritma *stemming* Nazief dan Adriani dikembangkan berdasarkan aturan morfologi Bahasa Indonesia yang mengelompokkan imbuhan menjadi awalan (*prefix*), sisipan (*infix*), akhiran (*suffix*) dan gabungan awalan akhiran (*confixes*). Algoritma ini menggunakan kamus kata dasar dan mendukung *recoding*, yakni penyusunan kembali kata-kata yang mengalami proses *stemming* berlebih (Nazief 2000). Aturan morfologi Bahasa Indonesia mengelompokkan imbuhan ke dalam beberapa kategori sebagai berikut:

1. *Inflection suffixes* yakni kelompok akhiran yang tidak merubah bentuk kata dasar. Sebagai contoh, kata "duduk" yang diberikan akhiran "-lah" akan menjadi "duduklah". Kelompok ini dapat dibagi menjadi dua:
   a. *Particle* (P) atau partikel yakni termaksud di dalamnya "-lah","kah", "tah" dan "pun".
   b. *Possessive pronoun* (PP) atau kata genti kepunyaan, termaksud di dalamnya "-ku","-mu" dan "-nya".
2. *Derivation suffixes* (DS) yakni kumpulan akhiran asli Bahasa Indonesia yang secara langsung ditambahkan pada kata dasar yaitu akhiran "-i", "-kan", dan "-an".
3. *Derivation prefixes* (DP) yakni kumpulan awalan yang dapat langsung diberikan pada kata dasar murni, atau pada kata dasar yang sudah mendapatkan penambahan sampai dengan 2 awalan termaksud di dalamnya adalah:
   a. Awalan yang dapat bermorfologi ("me-", "be-", "pe-" dan "te").
   b. Awalan yang tidak bermorfologi ("di-", "ke-" dan "se-").

Berdasarkan pengklasifikasi imbuhan-imbuhan di atas, maka bentuk kata berimbuhan dalam

$$\left[DP + \left[DP + [DP]\right]\right] KATA\ DASAR\ \left[[+DS] + PP\right]$$

Keterangan :
DP : *Derivation prefixes*
DS : *Derivation suffixes*
PP : *Possessive pronoun*

Dengan model Bahasa Indonesia di atas serta aturan-aturan dasar morfologi Bahasa Indonesia, aturan yang digunakan dalam proses algoritma Nazief dan Adriani sebagai berikut:

1. Tidak semua kombinasi awalan dan akhiran diperbolehkan. Kombinasi-kombinasi imbuhan yang tidak diperbolehkan, yaitu “be-i”, “ke-i”, “ke-kan”, “me-an”, “se-i”, “se-kan” dan “te-an”.
2. Penggunaan imbuhan yang sama secara berulang tidak diperkenankan.
3. Jika suatu kata hanya terdiri dari satu atau dua huruf, maka proses tidak dilakukan.
4. Penambahan suatu awalan tertentu dapat mengubah bentuk asli kata dasar, ataupun awalan yang telah diberikan sebelumnya pada kata dasar bersangkutan. Sebagai contoh, awalan “me-“ dapat berubah menjadi “meng-“, “men-“, “meny-“, dan “mem-“. Oleh karena itu diperlukan suatu aturan yang mampu mengatasi masalah morfologi ini.

Algoritma Nazief dan Adriani memiliki tahap-tahap sebagai berikut:

1. Cari kata dalam kamus jika ditemukan maka diasumsikan bahwa kata tersebut adalah kata dasar. Algoritma berhenti. Jika tidak ditemukan maka lakukan langkah 2.
2. Hilangkan *inflectional suffixes* bila ada. Dimulai dari *inflectional particle* (“-lah”, “-kah”, “-tah” dan “-pun”), kemudian *possessive pronoun* (“-ku”, “-mu” dan “-nya”). Cari kata pada kamus jikaditemukan algoritma berhenti, jika kata tidak tidak ditemukan dalam kamus lakukan langkah 3.
3. Hilangkan *derivation suffixes* (“-an”, “-i” dan “-kan”). Jika akhiran “-an” dihapus dan ditemukan akhiran “-k”, maka akhiran “-k” dihapus. Cari kata pada kamus jika ditemukan algoritma berhenti, jika kata tidak tidak ditemukan maka lakukan langkah 4.
4. Pada langkah 4 terdapat tiga iterasi.
   a. Iterasi berhenti jika:
      1) Ditemukannya kombinasi awalan yang tidak diizikan berdasarkan awalan.

Tabel 1. Kombinasi awalan dan akhiran yang tidak diizinkan

Awalan	Akhiran yang tidak diizinkan
be-	-i
di-	-an
ke-	-i, -kan
me	-an
se	-i,kan

      2) Awalan yang dideteksi saat ini sama dengan awalan yang dihilangkan sebelumnya.
      3) Tiga awalan telah dihilangkan.
   b. Identifikasikan tipe awalan dan hilangkan. Awalan terdiri dari dua tipe:
      1) Standar (“di-“, “ke-“, “se-”) yang dapat langsung dihilangkan dari kata.
      2) Kompleks (“me-“, “be-“, “pe-“, “te”) adalah tipe-tipe awalan yang dapat bermorfologi sesuai kata dasar yang mengikutinya. Oleh karena itu dibutuhkan aturan pada tabel berikut untuk mendapatkan hasil pemenggalan yang tepat.
   c. Cari kata yang telah dihilangkan awalannya. Apabila tidak ditemukan, maka langkah 4 diulang kembali. Apabila ditemukan, maka algoritma berhenti.

Setelah desain algoritma, dilakukan implementasi dan uji coba algoritma. Data untuk uji coba diambil dari kalimat-kalimat yang ditemukan pada portal berita daring. Berita dari portal daring diunduh dan disimpan otomatis. Kemudian dilakukan proses pemisahan antar kalimat dengan mendeteksi tanda baca titik (.), tanda seru (!), dan tanda tanya (?). Setelah setiap kalimat terpisah, dilakukan pemisahan tiap kata dengan mendeteksi spasi dan tanda koma (,). Kata yang terbentuk ini akan dilakukan proses stemming. Hasil implementasi algoritma dan uji coba ditunjukkan di Tabel 2. Empat puluh kata diuji coba untuk menguji kinerja algoritma infix-stripping. Keempat puluh kata tersebut mengandung sisipan el, em, dan er. Kemudian, hasil stemming dievaluasi oleh tiga orang validator Bahasa Indonesia. Validator diminta mengisi angka 0 jika hasil stemming tidak ditemukan pada Bahasa Indonesia baku dan angka 1 jika hasil stemming termasuk kata Bahasa Indonesia baku.

Tabel 2. Hasil Uji Coba Algoritma *Infix-stripping*

No	Kata Asli	Hasil *stemming*	Penilaian Validator		
			1	2	3
1	seruling	suling	1	1	1
2	kerudung	kudung	1	1	1
3	serabut	sabut	1	1	1
4	terperanjat	panjat	1	1	1
5	gerigi	gigi	1	1	1
6	genderang	gendang	1	1	1
7	reruntuhan	runtuh	1	1	1
8	reramuan	ramu	1	1	1
9	telunjuk	tunjuk	1	1	1
10	gelegar	gegar	1	1	1
11	telapak	tapak	1	1	1
12	penyelidikan	sidik	1	1	1
13	gelembung	gembung	1	1	1
14	jelajah	jajah	1	1	1
15	pelatuk	patuk	1	1	1
16	leluhur	luhur	1	1	1
17	melaju	maju	1	1	1
18	telangkup	tangkap	1	1	1
19	gemerlap	gerlap	1	1	1
20	gemetar	getar	1	1	1
21	gemilang	gilang	1	1	1
22	gemuruh	guruh	1	1	1
23	gemilap	gilap	1	1	1
24	kemelut	kelut	1	1	1
25	kemilap	kilap	1	1	1
26	kemilau	kilau	1	1	1
27	kemuning	kuning	1	1	1
28	semerbak	serbak	1	1	1
29	seminar	sinar	1	1	1
30	temali	tali	1	1	1
31	temaram	taram	1	1	1
32	temurun	turun	1	1	1
33	kemuncup	kuncup	1	1	1
34	jemari	jari	1	1	1
35	semilir	silir	1	1	1
36	gementar	gentar	1	1	1
37	gemertak	gertak	1	1	1
38	cemerlang	cerlang	1	1	1
39	peranjat	panjat	1	1	1
40	gemulung	gulung	1	1	1

ANALISIS HASIL PENELITIAN

Tabel 2 adalah hasil uji coba algortima *infix-stripping*. Berdasarkan hasil uji coba tersebut, dapat dilakukan analisa mengenai kata masukan yang dijadikan data uji coba, hasil penilaian oleh validator, dan kesimpulan algoritma infix-stripping yang telah diimplementasikan. Kata yang digunakan sebagai masukan uji coba berjumlah empat puluh kata dan semuanya mengandung kata sisipan. Beberapa kata mengandung imbuhan. Hasil stemming menunjukkan bahwa kata yang bersisipan dapat diubah menjadi kata dasar dengan menghilangkan sisipan el, em, atau er. Sedangkan kata yang masih berimbuhan, akan dihilangkan imbuhannya terlebih dahulu lalu dihilangkan sisipannya.

Tiga orang pakar Bahasa Indonesia diminta untuk memvalidasi hasil algoritma infix-stripping. Hasil validasi validator menunjukkan bahwa kata yang dihasilkan dari proses stemming merupakan kata baku Bahasa Indonesia. Validator 1 menilai seluruh (100%) kata dasar dari hasil stemming adalah kata dasar baku yang ditemukan dalam Kamus Besar Bahasa Indonesia. Validator 2 menilai seluruh (100%) kata dasar dari hasil stemming adalah kata dasar baku yang ditemukan dalam Kamus Besar Bahasa Indonesia. Demikian pula dengan validator 3, seluruh (100%) kata dasar dari hasil stemming adalah kata dasar baku yang ditemukan dalam Kamus Besar Bahasa Indonesia. Grafik yang menunjukkan hasil penilaian validator ditunjukkan di Gambar 2.

Gambar 2. Grafik penilaian validator Bahasa Indonesia

KESIMPULAN

Penelitian ini mengajukan ide untuk memperbaiki algoritma Nazief dan Adriani dan algoritma confix-stripping pada proses stemming Bahasa Indonesia dengan menambahkan fitur pelesapan sisipan. Beberapa kata bersisipan tanpa imbuhan dan berimbuhan diuji coba untuk mengetahui kinerja algoritma yang diajukan. Hasil penilaian yang dilakukan oleh ahli Bahasa Indonesia terhadap hasil stemming menunjukkan bahwa algoritma yang diajukan mampu mengembalikan kata sisipan ke bentuk aslinya. Penelitian selanjutnya yang dapat dikembangkan sebagai tindak lanjut dari studi ini adalah membandingkan hasil algoritma infix-stripping dengan algoritma Nazief dan Adriani dan algoritma confix-stripping.

DAFTAR PUSTAKA

Adriani, Mirna, Jelita Asian, Bobby Nazief, and Hugh E Williams. 2007. “Stemming Indonesian : A Confi X-Stripping Approach.” ACM Transactions on Asian Language Information Processing (TALIP) 6 (4):1–33.

Nazief, Bobby. 2000. “Development of Computational Linguistics Research: A Challenge for Indonesia.” Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL ’00, 1–2.

Setiawan, Reina, Aditya Kurniawan, Widodo Budiharto, Iman Herwidiana Kartowisastro, and Harjanto Prabowo. 2016. “Flexible Affix Classification for Stemming Indonesian

Language." In 2016 13th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, ECTI-CON 2016.

Wibowo, Julianto. 2016. "Pada Kalimat Bahasa Indonesia Dengan Algoritma Stemming." Jurnal Riset Komputer (JURIKOM) 3 (5):346–50.

Widayanto, Hari, and Arief Fatchul Huda. 2017. "Comparison Nazief Adriani And CS Stemmer Algorithm For Stemm Real Data." E-Proceeding of Engineering 4 (3):5215–22.

Widjaja, Marsel, and Seng Hansun. 2015. "Implementation of Porter's Modified Stemming Algorithm in an Indonesian Word Error Detection Plugin Application." International Journal of Technology 6 (2):139–50.

Winarti, Titin, Jati Kerami, and Sunny Arief. 2017. "Determining Term on Text Document Clustering Using Algorithm of Enhanced Confix Striping Stemming." International Journal of Computer Application 157 (9):6.