Pengembangan Algoritma Stemmer Bilingual Bali-Indonesia Dengan Rule-Base

Gusti Ngurah Mega Nata, I Gusti Ngurah Nyoman Bagiarta, I Putu Ramayasa, I Made Ari Santosa
Fakultas Bisnis dan Vokasi, Fakultas Informatika dan komputer
ITB STIKOM Bali
Denpasar, Indonesia
e-mail: mega@stikom-bali.ac.id,

Abstrak

Bahasa Bali sekarang sudah banyak kata serapan dari Indonesia dan kadang dicampur dalam penulisannya dengan Bahasa Indonesia atau Bahasa asing lainnya. Dokumen bilingual Bali-Indonesia akan menjadi masalah pada saat proses temu balik informasi / information retrival menggunakan mesin pencari. System Information retrival yang berkembang sampai penelitian ini dilakukan belum ditemukan information retrival bilingual Bali-Indonesia. Dalam information retrival pada dokumen bilingual khusunya Bali-Indonesia akan menemukan kesulitan dalam tahap text preprocessor yaitu stemming. Permasalahan tersebut karna Selama ini proses text preprocessing untuk Bahasa Bali hanya khusus untuk Bahasa Bali saja. Maka, dari permasalahan tersebut perlu dikembangkan sebuah algoritma stemmer untuk bilingual Bahasa Bali-Indonesia. Bahasa Bali dan Bahasa Indonesia jika dilihat dari struktur morfologi afiks yang dimiliki hampir sama yaitu terdapatnya prefiks, supfiks, infiks, konfiks, simulfiks dan kombinasi afiks hanya saja jumlah dan karakter pembentuknya yang berbeda. Maka, pada paper ini pengembangkan stemmer Bilingual Bali-Indonesia dengan cara hybrid afiks dari kedua Bahasa tersebut kemudian proses stemmer menggunakan teknik rule-base yaitu dengan memotong afiks dari setiap kata algoritma kedua yaitu dengan melakukan dua kali stemmer yaitu Bali dan Indonesia. Hasil analisis Algoritma pengabungan prefixs dan surfixs secara konsep memiliki kecepatan yang labih baik dibandingkan dengan penggabungan algoritma porter stemmer bahasa Indonesia dengan algoritma stemmber bahasa Bali.
Kata kunci: *Billingual Bahasa Bali. Stemmer, hybrid afiks*

Abstract

The Balinese language now has many absorption words from Indonesia and is sometimes mixed in writing with Indonesian or other foreign languages. Bilingual Balinese-Indonesian documents will become a problem during the information retrieval process using search engines. The Information retrival system that was developed until this research was conducted has not yet found a Balinese-Indonesian bilingual information retrieval. In information retrieval on bilingual documents, especially Bali-Indonesia, there will be difficulties in the text preprocessor stage, namely stemming. This problem is because so far the text preprocessing process for Balinese has only been specifically for Balinese. So, from these problems it is necessary to develop a stemmer algorithm for bilingual Balinese-Indonesian. Balinese and Indonesian, when viewed from the morphological structure of the affixes, are almost the same, namely there are prefixes, suffixes, infixes, confixes, simulfixes and combinations of affixes, but the number and characters of their constituents are different. So in this study the development of a Balinese-Indonesian bilingual stemmer by means of hybrid affixes from the two languages was then processed using a rule-base technique, namely by cutting off the affixes of each word. The results of the analysis of the combining prefixes and surfixs algorithms conceptually have better speed compared to combining Indonesian porter stemmer algorithms with Balinese stemmber algorithms.
Keywords*: Bilingual Balinese-Indonesian. Stemmer, hybrid affix*

1. Pendahuluan

Bahasa Bali sekarang sudah banyak kata serapan dari Indonesia dan kadang dicampur dalam penulisannya dengan Bahasa Indonesia. Penggunaan kedua Bahasa tersebut sudah menjadi hal yang biasa dalam komunikasi di media social, penulisan surat edaran, lirik lagu atau dalam komunikasi lainnya. Hal tersebut tentunya sangat sulit dihindari karena pergaulan remaja Bali yang sudah mulai lintas budaya dan bahasa [1]. Permasalahan penggunaan bilingual tersebut dalam komunikasi langsung tentunya tidak perlu lagi diperdebatkan, yang penting lawan bicara kita saling mengerti itu sudah cukup. Namun, jika dokumen bilingual Bali-Indonesia tentunya akan menjadi masalah pada saat proses temu balik informasi / information retrival menggunakan mesin pencari [2].

System Information retrival yang berkembang sampai penelitian ini dilakukan belum ditemukan *information retrival* bilingual Bali-Indonesia. Dalam information retrival pada dokumen bilingual

khusunya Bali-Indonesia akan menemukan kesulitan dalam tahap *text preprocessing*. Permasalahan tersebut karna Selama ini proses *text preprocessing* untuk Bahasa Bali hanya khusus untuk Bahasa Bali saja seperti pada paper [3] atau pada paper [4]. Sehingga kedua teknik dari kedua paper tersebut belum mampu melakukan stemming dokumen bilingual yaitu Bali-Indonesia. Maka, dari permasalahan tersebut perlu dikembangkan sebuah algoritma stemmer untuk bilingual bahasa Bali-Indonesia dan juga pemanfaatanya untuk information retrival. bahasa Bali dan bahasa Indonesia jika dilihat dari struktur afiks yang dimiliki hampir sama yaitu terdapatnya prefiks, supfiks, infiks, konfiks, simulfiks dan kombinasi afiks [4] hanya saja jumlah dan karakter pembentuknya yang berbeda. Maka pada paper ini akan pengembangkan stemmer Bilingual Bali-Indonesia dengan cara *hybrid* afik dan kata dasar dari kedua Bahasa tersebut kemudian proses stemmer menggunakan teknik rule-base yaitu dengan memotong afiks dari setiap kata dan mencocokannya pada list kata dasar model algoritma kedua yaitu dengan mengabungkan algoritma stemmber bahasa Bali dengan stemmber bahasa Indonesia.

2. Metode Penelitian

Stemming merupakan salah satu proses preprocessing dokumen teks, untuk proses data mining [5]. Tujuan dari stemmer adalah untuk mendapatkan kata dasar dari sebuah kata dalam teks. Sebelum proses stemmer terdapat beberapa tahapan yaitu pengumpulan dokumen teks, parsing, stopword removal setelah itu baru stemming.

2.1. Pengumpulan Data teks billingual Bali-Indonesia

Data yang digunakan untuk uji coba adalah data teks yang berisi bahasa Bali dan bahasa Indonesia. Data uji terlebih dahulu di cek agar tidak ada kesalahan ketik atau singkatan.

2.2. *Parsing / Tokenizing*

Teks adalah data *unstructured* yang harus dirubah dahulu menjadi terstuktur sebelum dianalis lebih lanjut. Bentuk *unstructured* dari teks adalah kata – kata yang terurut (*sequence*) dan menyambung dalam sebuah dokumen. Proses perubahan bentuk teks menjadi data terstruktur disebut dengan teks *pre-processing*. Teks processing diawali dengan memotong setiap kata yang ada dalam teks tersebut menjadi perkata. Process pemotongan teks menjadi kata – kata terpisah disebut *parsing / tokenizing*.

2.3. *Stopword* removal

Akibat dari *tokenizing* setiap kata berdiri sendiri dan tidak memiliki arti yang relevan untuk menentukan ciri dari dokumen yang di *tokenizing* seperti "*ini, itu, adalah, dan, atau*" dan bayak lagi kata – kata sejenis. Kata – kata yang tidak memiliki arti yang relevan tersebut disebut *stop word.* Kumpulan dari stop word disebut stop list dan proses untuk menghapus stop word dalam dokumen disebut *stopword removal.* stopword list dalam paper ini digabungan dari stop list bahasa Bali dan stop list bahasa Indonesia untuk algoritma hybrid afik.

2.4. *Stemming*

Stemming adalah proses pemetaan dan penguraian berbagai bentuk dari suatu kata menjadi bentuk kata dasarnya [6]. Proses *stemming* untuk setiap Bahasa berbeda-beda misal, proses *stemming* Bahasa Inggris dengan Bahasa Indonesia tentunya berbeda karena perbedaan pembentukan dan perubahan kata menjadi bentuk kata lain [2]. Namun Bahasa Indonesia dengan bahasa Bali memiliki kesamaan morfologi tetapi memiliki *prefixes, suffixes, infexes* dan *confixes* yang tidak sama. Berikut adalah arti dan contoh dari imbuhan [5]. Pada dasarnya Teknik stemmer yang telah dikembangkan dari penelitian yang sudah dilakukan yaitu dapat dibagi menjadi dua

1. Stemmer Menggunakan kamus kata dasar, seperti penelitian yang telah dilakukan oleh Nazief dan Adriani dari Universitas Indonesia pada tahun 1996,2007.
2. Stemmber tidak menggunakan kamus kata dasar, seperti yang telah dilakukan oleh Fadillah Z. Tala pada tahun 2002 [8].

2.4.1 Porter Stemming Bahasa Indonesia

Porter Stemming Bahasa Indonesia merupakan algoritma untuk pemetaan dan penguraian berbagai bentuk kata menjadi bentuk kata dasarnya. Proses porter stemming untuk setiap Bahasa berbeda dengan Bahasa yang lain misal, proses porter stemming Bahasa Inggris dengan Bahasa Indonesia tentunya berbeda karena perbedaan pembentukan dan perubahan kata menjadi bentuk kata lain [9].

Porter Stemmer for Bahasa Indonesia dikembangkan oleh Fadillah Z. Tala pada tahun 2002. Implementasi Porter Stemmer for Bahasa Indonesia berdasarkan English Porter Stemmer yang dikembangkan oleh W.B. Frakes pada tahun 1992. Karena bahasa Inggris datang dari kelas yang berbeda, beberapa modifikasi telah dilakukan untuk membuat Algoritma Porter dapat digunakan sesuai dengan bahasa Indonesia. Meskipun terdapat infiks dalam bahasa indonesia dan jumlahnya sedikit, dalam penelitian yang dilakukan pada paper [8], kata yang memilki infiks akan dianggap kata itu sendiri.

Algoritma porter stemmer Bahasa Indonesia dalam mencari kata dasar tidak menggunakan kamus kata dasar. Teknik algoritma porter Bahasa Indonesia dalam mencari kata dasar yaitu dengan melihat dan mengapus Affixes dari Bahasa Indonesia. sehingga algoritma porter stemmer memiliki kecepatan yang lebih dibandingkan algoritma stemmer yang menggunakan kamus kata dasar seperti yang dikembangkan oleh Nazief dan Adriani dari Universitas Indonesia pada tahun 1996,2007.

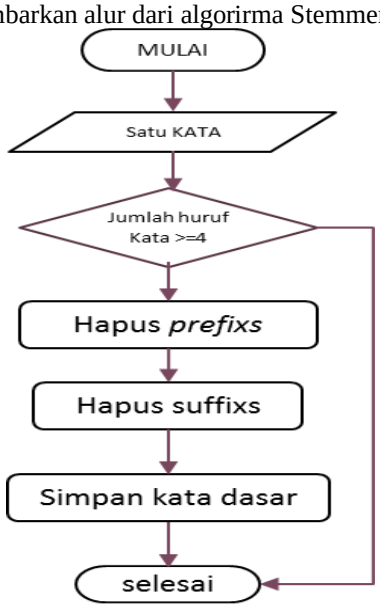
2.4.2 Stemmer bahasa Bali

Proses *stemming* Bahasa Bali *sor-singgih* mengadopsi cara kerja dari algoritma porter *stemmer* Bahasa Indonesia [3]. Proses *stemmer* menggunakan cara kerja seperti algoritma porter yaitu dengan memotong imbuhan (*prefixs)*. Namun imbuhan yang digunakan hanya *prefixes* dan *suffixes.* Bahasa Bali tidak memiliki partikel baku yang secara langsung menempel pada kata dasarnya [1], jadi pada stemmer bahasa bali tidak ada proses hapus *partikel* seperti pada algoritma porter. Begitu juga dengan *possessive pronoun*, Bahasa Bali untuk *prossesive pronoun* tidak secara langsung menempel pada kata kerja. Maka algoritma ini hanya mengapus awalan (*prefixs*) dan akhiran (*suffixs*).

Cara kerja algoritma *stemmer* sebagai berikut:

a. Jika jumlah huruf dalam kata minimal 4 huruf maka proses lebih lanjut, jika kurang dari 4 proses stemmer berhenti.
b. Hapus Awalan (*prefixs*).
c. Hapus akhiran (*Suffixs*) dari kata yang dimasukan
d. Simpan kata dasar pada database
e. Kata dasar sor-singgih hasil *stemming* kemudian dicarikan padanannya dengan Bahasa Indonesia agar tingkatan sor-singgih bahasa Bali di jadikan satu tingkat. Proses pencarian padanan kata menggunakan daftar padanan kata dasar.

Pada gambar 1 berikut digambarkan alur dari algorirma Stemmer Bahasa Bali.



Gambar 1. flowchart stemmer Bahasa Bali

3. Hasil dan Pembahasan

3.1. Pengabungan Afek dan suffix bali-indonesia

Pengabungan Affixs dan suffix dari bahasa Bali dengan bahasa Indonesia didasari dari kesamaan penyusun kata dari kedua bahasa ini. Hanya saja elemen pada setiap imbuhan tidak semuanya sama. Berikut adalah penggabungan Affix dan suffix bahasa Indonesia dengan bahasa Bali.

Tabel 1. Pengabungan awalan (prefixes)

Prefixes	Bahasa
Mem	Indonesia
per	Indonesia
Ma	Bali
Pa	Bali
Ka	Bali
Di	Bali
Sa	Bali
Ny	Bali
M	Bali
N	Bali
Ng	Bali

Tabel 2. Pengabungan Akhiran (suffixes)

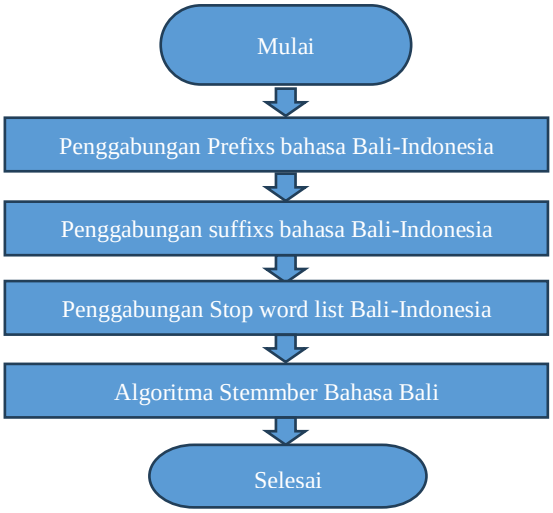
Suffixes	Bahasa
-e, -ne, -an, -ang, -n, -in, -ing	Bali
-Lah, Kah, Ku, Mu, Nya, Tah, Pun, Mu, Nya, Kan, Mem, per	Indonesia

3.2. Desing Algoritma Stemmer billingual Bali-Indonesia

Algoritma Billingan Bali-Indonesia dengan pengabungan Afiks

Algoritma pertama yaitu dengan mengabungkan masing – masing dari prefiks dan suffixs Bali-Indonesia. Proses pertama yaitu penggabungan prefixs, proses kedua yaitu penggabungan suffixs, ketiga penggabunganstop word list dan terahir baru proses stemming menggunakan algoritma stemmber bahasa bali. Berikut adalah urutan algoritma stemmer dengan cara penggabungan afiks Bahasa Bali dengan Bahasa Indonesia.

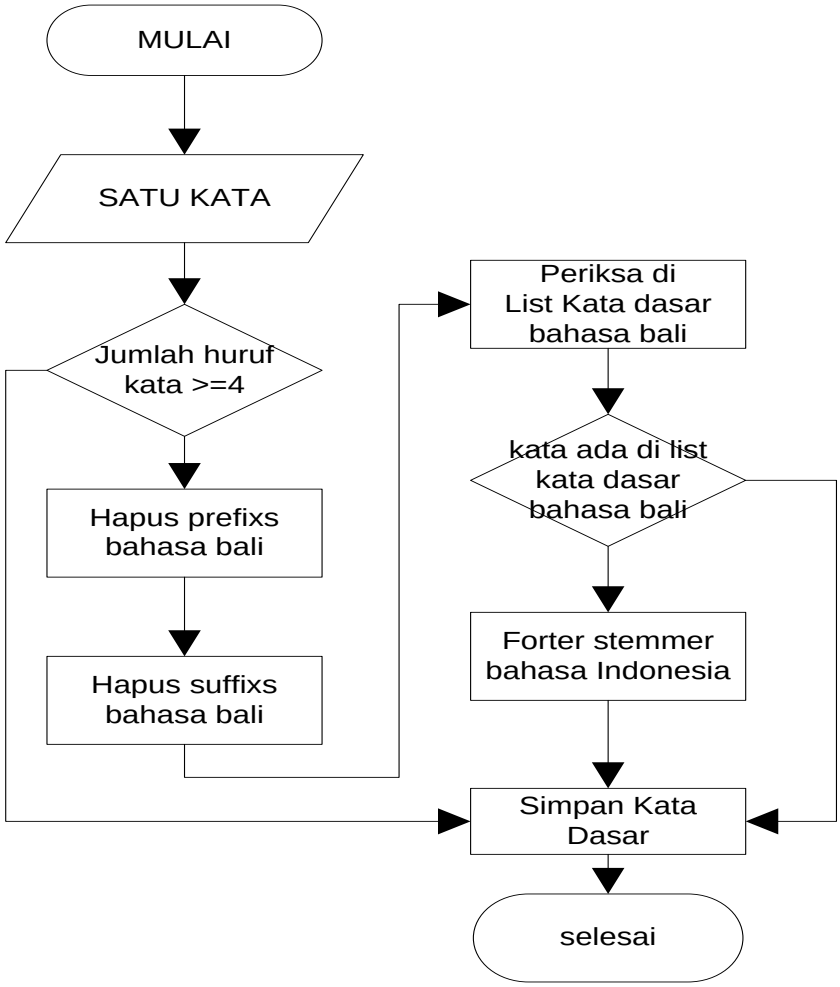
a. Pengabungan *prefix* Bahasa bali dengan *prefix* Bahasa Indonesia dilakukan diawal stemmer.
b. Pengabungan *suffixs* Bahasa bali dengan *suffixs* Bahasa Indonesia
c. Pengabungan stop word list Bali-Indonesia
d. Porter stemmer Bahasa Bali menggunakan algoritma pada gambar 1.



Gambar 2. Algoritma Stemmer Billingual dengan pengabungan afiks Bali-Indonesia

Algoritma Billingual Bali-Indonesia dengan Pengabungan Proses Stemmer

Algoritma bilingual Bali-Indonesia dengan pengabungan proses stemmer memiliki proses seperti pada gambar berikut:



Gambar 3. Algoritma Stemmer Billingual Bali-Indonesia dengan Pengabungan proses

Pembentukan dari algoritma ini dimulai dari stemmer Bahasa Bali kemudian pengecekan pada list kata dasar dan kemudian baru stemmer Bahasa Indonesia.

4. Kesimpulan

Kesimpulan yang dapat diambil yaitu.

a. Pengujian yang dilakukan untuk mengukur akurasi dan kecepatan dalam proses. Algoritma pengabungan *prefixs* dan *surfixs* secara konsep memiliki kecepatan yang labih baik.
b. Pengembangan algoritma stemmer bilingual Bali-Indonesia dikembangkan menggunakan metode Rule-Base karena Bahasa bali dan Bahasa Indonesia memiliki morfologi yang sama.
c. Jumlah kata dasar Bahasa Bali akan mempengaruhi akurasi pada algoritma ke dua yaitu pengabungan stemmer Bali-Indonesia.

Daftar Pustaka

[1] I. B. Paramita, "Kontemplasi: Komunikasi, Etika Dan Pengetahuan Dalam Bahasa Bali," *J. Comun.*, pp. 191–200, 2020.

[2] G. N. M. Nata, "Information Retrival untuk Pencarian Dokumen Tugas Akhir Menggunakan Sequential Pattern Mining," *Semin. Multimed. \& Artif. ...*, no. 84, pp. 81–86, 2018, [Online]. Available: http://papersmai.mercubuana-yogya.ac.id/index.php/smai/article/view/13

[3] G. N. M. Nata and P. P. Yudiastra, "Stemming teks sor-singgih Bahasa Bali," *E-Proceedings KNS&I STIKOM Bali*, pp. 608–612, 2017, [Online]. Available: http://knsi.stikom-bali.ac.id/index.php/eproceedings/article/view/111

[4] M. A. P. Subali and S. Rochimah, "A new model for measuring the complexity of SQL commands," *Proc. 2018 10th Int. Conf. Inf. Technol. Electr. Eng. Smart Technol. Better Soc. ICITEE 2018*, pp. 1–5, 2018, doi: 10.1109/ICITEED.2018.8534782.

[5] G. Ngurah, M. Nata, and P. P. Yudiastra, "Knowledge discovery pada email box sebagai penunjang email marketing knowledge discovery in the email box for support email marketing," *J. Sist. dan Inform.*, pp. 26–37, 2017.

[6] D. Rahmawati, G. A. P. Saptawati, and Y. Widyani, "Document Clustering using Sequential Pattern(SP)," *2015 Int. Conf. Data Softw. Eng. (ICoDSE 2015)*, pp. 98–102, 2015, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7436979

[7] A. A. Magriyanti, "Analisis Pengembangan Algoritma Porter Stemming Dalam Bahasa Indonesia," *Sekol. Tinggi Elektron. dan Komput. PAT* , vol. 1, no. 1, pp. 1–5, 2018, [Online]. Available: http://kompas.com

[8] F. Z. Tala, "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia," *M.Sc. Thesis, Append. D*, vol. pp, pp. 39–46, 2003.

[9] G. Ngurah, M. Nata, P. Yudiastra,) Stmik, S. Bali, and J. Raya Puputan, "Preprocessing Text Mining Pada Email Box Berbahasa Indonesia," *E-Proceedings KNS&I STIKOM Bali*, pp. 479–483, 2017, [Online]. Available: http://www.knsi.stikom-bali.ac.id/index.php/eproceedings/article/view/88