ISOLATED-WORD ERROR CORRECTION PADA TEKS BERBAHASA INDONESIA

Skripsi

oleh
AMADEA KRISTINA BUDIMAN
71110057

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2016

ISOLATED-WORD ERROR CORRECTION PADA TEKS BERBAHASA INDONESIA

Skripsi

Diajukan kepada Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana
Sebagai Salah Satu Syarat dalam Memperoleh Gelar
Sarjana Komputer

Disusun oleh

AMADEA KRISTINA BUDIMAN
71110057

PROGRAM STUDI TEKNIK INFORMATIKA FAKULTAS TEKNOLOGI INFORMASI
UNIVERSITAS KRISTEN DUTA WACANA
2016

PERNYATAAN KEASLIAN SKRIPSI

Saya menyatakan dengan sesungguhnya bahwa skripsi dengan judul:

ISOLATED-WORD ERROR CORRECTION PADA TEKS BERBAHASA INDONESIA

yang saya kerjakan untuk melengkapi sebagian persyaratan menjadi Sarjana Komputer pada pendidikan Sarjana Program Studi Teknik Informatika Fakultas Teknologi Informasi Universitas Kristen Duta Wacana, bukan merupakan tiruan atau duplikasi dari skripsi kesarjanaan di lingkungan Universitas Kristen Duta Wacana maupun di Perguruan Tinggi atau instansi manapun, kecuali bagian yang sumber informasinya dicantumkan sebagaimana mestinya.

Jika dikemudian hari didapati bahwa hasil skripsi ini adalah hasil plagiasi atau tiruan dari skripsi lain, saya bersedia dikenai sanksi yakni pencabutan gelar kesarjanaan saya.

Yogyakarta, 4 Januari 2016

AMADEA KRISTINA BUDIMAN
71110057

HALAMAN PERSETUJUAN

Judul Skripsi	:	ISOLATED-WORD ERROR CORRECTION PADA TEKS BERBAHASA INDONESIA
Nama Mahasiswa	:	AMADEA KRISTINA BUDIMAN
N I M	:	71110057
Matakuliah	:	Skripsi (Tugas Akhir)
Kode	:	TIW276
Semester	:	Gasal
Tahun Akademik	:	2015/2016

Telah diperiksa dan disetujui di
Yogyakarta,
Pada tanggal 4 Januari 2016

Dosen Pembimbing I	Dosen Pembimbing II
Gloria Virginia, S.Kom., MAI, Ph.D.	Budi Susanto, SKom.,M.T.

HALAMAN PENGESAHAN

ISOLATED-WORD ERROR CORRECTION PADA TEKS BERBAHASA INDONESIA

Oleh: AMADEA KRISTINA BUDIMAN / 71110057

Dipertahankan di depan Dewan Penguji Skripsi
Program Studi Teknik Informatika Fakultas Teknologi Informasi
Universitas Kristen Duta Wacana - Yogyakarta
Dan dinyatakan diterima untuk memenuhi salah satu syarat memperoleh gelar
Sarjana Komputer
pada tanggal 10 Desember 2015

Yogyakarta, 4 Januari 2016
Mengesahkan,

Dewan Penguji:

1. Gloria Virginia, S.Kom., MAI, Ph.D.
2. Budi Susanto, SKom.,M.T.
3. R. Gunawan Santosa, Drs. M.Si.
4. Hendro Setiadi, M.Eng

Dekan	Ketua Program Studi
(Budi Susanto, S.Kom., M.T.)	(Gloria Virginia, Ph.D.)

UCAPAN TERIMA KASIH

Puji syukur penulis naikkan ke hadirat Tuhan Yang Maha Esa yang telah melimpahkan anugerah dan rahmat-Nya, sehingga penulis dapat menyelesaikan program dan laporan tugas akhir berjudul "Isolated-Word Error Correction Pada Teks Berbahasa Indonesia" ini dengan baik dan tepat waktu.

Dalam menyelesaikan penelitian ini, penulis menyadari banyak menerima masukan dan saran dari berbagai pihak, baik secara langsung dan secara tidak langsung. Penulis juga telah mendapatkan banyak dukungan dari beberapa pihak. Oleh karena itu, penulis ingin mengucapkan terima kasih kepada:

1. Keluarga terkasih yang senantiasa mendoakan, memberi dukungan dan motivasi kepada penulis selama ini,
2. Ibu Gloria Virginia, S.Kom., MAI, Ph.D. dan Bapak Budi Susanto, S.Kom., M.T. yang telah mendukung, membimbing, dan memberi berbagai masukan kepada penulis terkait pengerjaan tugas akhir ini,
3. Teman-teman yang telah memberikan masukan dan dorongan semangat kepada penulis,
4. Pihak-pihak yang tidak dapat disebutkan satu per satu yang telah membantu penulis dalam menyelesaikan penelitian tugas akhir ini melalu berbagai hal.

Akhir kata, penulis ingin meminta maaf apabila terdapat kesalahan dalam penyusunan laporan tugas akhir ini. Terima kasih.

Yogyakarta, 4 Januari 2016

Amadea Kristina Budiman

ABSTRAKSI

ISOLATED-WORD ERROR CORRECTION PADA TEKS BERBAHASA INDONESIA

Data mentah yang terdapat dalam dokumen teks biasanya tidak terstruktur bentuknya sehingga tidak bisa digunakan untuk mendapatkan informasi secara baik. Oleh karena itu, diterapkan *text preprocessing* untuk mengolah suatu data mentah yang berupa teks. Salah satu penerapan *text preprocessing* adalah *spelling correction*. Dalam *spelling correction* terdapat teknik yang mengoreksi *term* tanpa memperhatikan konteks dari teks. Teknik tersebut bernama *isolated-word error correction*. Sebelum melakukan koreksi, diperlukan tahap *nonword error detection* yang berfungsi untuk memeriksa ejaan pada suatu *term.*

Dalam penelitian ini penulis meneliti penerapan *nonword error detection* dan penggunaan Dice *coefficient* untuk melakukan *isolated-word error correction* serta Damerau-Levenshtein *distance* dan *word frequency*. Penulis juga meneliti pengaruh teknik tokenisasi pada k-gram menggunakan kombinasi susunan karakter. Evaluasi dilakukan menggunakan nilai *precision*, *recall*, dan *f-measure* pada 100 dokumen teks yang diambil dari ICL-*corpus*. Evaluasi dilakukan terhadap proses deteksi kesalahan dan koreksi kesalahan. Pada evaluasi koreksi kesalahan, setiap kombinasi metode yang menggunakan koefisien Dice akan dievaluasi berdasarkan variasi nilai *threshold* Dice, yaitu sebesar 0.2, 0.4, 0.6, dan 0.8.

Pada evaluasi deteksi kesalahan, didapat nilai *f-measure* sebesar 0.86517. Sedangkan kombinasi metode yang dapat memberikan hasil koreksi terbaik adalah bigram kombinasi yang menggunakan koefisien Dice, Damerau-Levenshtein, dan frekuensi *term* pada nilai *threshold* Dice sebesar 0.2. Nilai *f-measure* dari metode tersebut adalah 0.50112.

Kata kunci: Damerau-Levenshtein, Dice *coefficient*, *isolated-word error correction*, *nonword error detection*, *word frequency*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

ABSTRAKSI

ISOLATED-WORD ERROR CORRECTION PADA TEKS BERBAHASA INDONESIA

Data mentah yang terdapat dalam dokumen teks biasanya tidak terstruktur bentuknya sehingga tidak bisa digunakan untuk mendapatkan informasi secara baik. Oleh karena itu, diterapkan *text preprocessing* untuk mengolah suatu data mentah yang berupa teks. Salah satu penerapan *text preprocessing* adalah *spelling correction*. Dalam *spelling correction* terdapat teknik yang mengoreksi *term* tanpa memperhatikan konteks dari teks. Teknik tersebut bernama *isolated-word error correction*. Sebelum melakukan koreksi, diperlukan tahap *nonword error detection* yang berfungsi untuk memeriksa ejaan pada suatu *term.*

Dalam penelitian ini penulis meneliti penerapan *nonword error detection* dan penggunaan Dice *coefficient* untuk melakukan *isolated-word error correction* serta Damerau-Levenshtein *distance* dan *word frequency*. Penulis juga meneliti pengaruh teknik tokenisasi pada k-gram menggunakan kombinasi susunan karakter. Evaluasi dilakukan menggunakan nilai *precision*, *recall*, dan *f-measure* pada 100 dokumen teks yang diambil dari ICL-*corpus*. Evaluasi dilakukan terhadap proses deteksi kesalahan dan koreksi kesalahan. Pada evaluasi koreksi kesalahan, setiap kombinasi metode yang menggunakan koefisien Dice akan dievaluasi berdasarkan variasi nilai *threshold* Dice, yaitu sebesar 0.2, 0.4, 0.6, dan 0.8.

Pada evaluasi deteksi kesalahan, didapat nilai *f-measure* sebesar 0.86517. Sedangkan kombinasi metode yang dapat memberikan hasil koreksi terbaik adalah bigram kombinasi yang menggunakan koefisien Dice, Damerau-Levenshtein, dan frekuensi *term* pada nilai *threshold* Dice sebesar 0.2. Nilai *f-measure* dari metode tersebut adalah 0.50112.

Kata kunci: Damerau-Levenshtein, Dice *coefficient*, *isolated-word error correction*, *nonword error detection*, *word frequency*

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Data mentah yang terdapat dalam dokumen teks biasanya memiliki bentuk yang tidak terstruktur sehingga tidak bisa digunakan untuk mendapatkan informasi secara baik. Oleh karena itu, digunakan *text preprocessing* untuk mengolah suatu data mentah yang berupa teks. Salah satu teknik dalam *text preprocessing* adalah *spelling correction*. Dalam *spelling correction* terdapat teknik yang mengoreksi *term* tanpa memperhatikan konteks dari teks. Teknik tersebut bernama *isolated-word error correction* yang berfungsi untuk mengoreksi *nonword error*. Meskipun sudah terdapat penelitian tentang *isolated-word error correction* pada *term* bahasa Indonesia, mayoritas penelitian-penelitian tersebut hanya menggunakan Levenshtein *distance*, seperti penelitian yang dilakukan oleh Luqman (2009), Dwitiyastuti *et al.* (2013), dan Atmajaya (2008).

Melihat masih besarnya peluang penelitian yang dapat dilakukan, penulis ingin melakukan penelitian implementasi *isolated-word error correction* pada teks berbahasa Indonesia. Sebelum melakukan *isolated-word error correction*, perlu dilakukan *nonword error detection* terlebih dahulu. *Error detection* merupakan tahap pemeriksaan ejaan pada suatu *term*. Pada tahap *nonword error detection*, penulis menggunakan beberapa kamus, yaitu kamus bahasa Indonesia, kamus bahasa Inggris, kamus akronim dan singkatan. Penulis juga menggunakan daftar *proper noun* dan daftar kata nonformal yang didapat dari ICL-*corpus*.

Algoritma yang penulis pilih untuk melakukan *isolated-word error correction* adalah salah satu algoritma berbasis k-gram, yaitu Dice *coefficient*. Pemilihan algoritma tersebut berdasarkan alasan penggunaan rumus yang sederhana sehingga tidak memerlukan waktu yang banyak untuk melakukan proses perhitungan. Pada teknik tokenisasi k-gram yang digunakan oleh Dice, penulis meneliti penggunaan k-gram dengan teknik *sliding window* dan k-gram

yang menggunakan kombinasi susunan karakter.

Selama proses pembangunan sistem, penulis mengamati bahwa Dice *coefficient* kurang baik jika digunakan sendirian untuk melakukan pengurutan *term* pada tahap *isolated-word error correction*, karenanya penulis ingin meneliti juga penggunaan algoritma Damerau-Levenshtein *distance* dan frekuensi kemunculan *term* (*word frequency*) guna meningkatkan performa sistem pada tahap *isolated-word error correction*.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah di atas, maka permasalahan yang akan diteliti meliputi :

a. Bagaimana performa *nonword error detection* pada sistem yang dibangun?
b. Bagaimana performa algoritma Dice *coefficient* dalam melakukan *isolated-word error correction*?
c. Bagaimana performa *isolated-word error correction* yang menggunakan k-gram yang dengan kombinasi susunan abjad dibandingkan dengan k-gram yang menggunakan teknik *sliding window*?
d. Apakah ada pengaruh terhadap performa pada *isolated-word error correction* ketika menggunakan algoritma Damerau-Levenshtein *distance* dan frekuensi kemunculan *term*?

1.3 Batasan Masalah

Penelitian yang akan dilakukan memiliki beberapa batasan sebagai berikut ini :

a. Sistem dibuat dalam bentuk aplikasi *desktop*.
b. Koreksi dilakukan terhadap dokumen teks tunggal dengan mayoritas penggunaan *term* adalah *term* berbahasa Indonesia.
c. Koreksi dilakukan tanpa memperhatikan konteks dari teks.
d. Konten dokumen uji diambil dari ICL-*corpus*. Dokumen yang diambil

berjumlah 100 dokumen sebagai perwakilan dari *corpus* tersebut.

e. *Gold-standard* dari 100 dokumen uji dibuat dengan mempertimbangkan penggunaan kata nonformal.
f. Daftar kata untuk kamus bahasa Indonesia diambil dari http://kateglo.com/?mod=dictionary. Kata yang diambil adalah kata yang bersumber dari Kamus Besar Bahasa Indonesia (KBBI) edisi ketiga.
g. Kamus bahasa Inggris diunduh dari http://downloads.sourceforge.net/wordlist/12dicts-5.0.zip. Daftar kata yang digunakan adalah daftar kata yang berada pada dokumen teks berjudul 2of12inf.txt dan 6of12.txt
h. Kamus akronim dan singkatan didapat dari lampiran KBBI edisi keempat yang dapat dilihat di https://www.scribd.com/doc/33910381/Daftar-singkatan-dan-akronim-Lampiran-KBBI-IV.
i. Daftar *proper noun* hanya berasal dari ICL-*corpus*.
j. Daftar kata nonformal hanya berasal dari kata-kata nonformal dalam 100 dokumen uji dari ICL-*corpus* yang dipilih oleh penulis.
k. Frekuensi kemunculan *term* didapat dari ICL-*corpus* dan Kompas-*corpus*.

1.4 Tujuan Penelitian

Tujuan utama dari penelitian ini adalah mengetahui performa koefisien Dice dalam melakukan *isolated-word error correction* serta untuk mengetahui pengaruh penggunaan k-gram kombinasi, algoritma Damerau-Levenshtein *distance*, dan frekuensi *term* dalam memberikan pembenaran *term*. Selain itu, penelitian ini juga bertujuan untuk mengetahui performa *nonword error detection* pada sistem yang dibangun.

1.5 Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah sebagai berikut:

a. Studi literatur / pustaka

   Merupakan tahap awal dari penelitian dan perancangan sistem. Pada tahap ini penulis melakukan peninjauan berbagai literatur/pustaka dalam bentuk teks buku, jurnal, dan *e-book* yang membahas tentang *text mining, data preprocessing, string matching, error detection, error correction*, dan *similarity measures*. Dari studi literatur tersebut, penulis memilih menggunakan *nonword error detection* dan algoritma Dice *coefficient*, Damerau-Levenshtein serta *word frequency* untuk melakukan *isolated-word error correction*.

b. Pengumpulan data

   Merupakan tahap untuk mengetahui kebutuhan sistem yang akan dibangun, seperti bahasa pemrograman yang akan digunakan untuk membangun sistem.

c. Pembangunan sistem

   Merupakan tahap untuk melakukan perancangan detail sistem dari data yang sudah didapatkan pada tahap pengumpulan data tahap pengerjaan sistem berdasarkan perancangan sistem yang telah dibuat. Kemudian dilakukan pembuatan sistem yang terdiri dari penulisan *source code* program, penelusuran kesalahan sistem, perbaikan kesalahan sistem sehingga sistem dapat memenuhi tujuan awal pembuatan sistem.

d. Evaluasi

   Tahap ini dilakukan dengan menggunakan dokumen uji dari ICL-*corpus*. Kemudian dari hasil pengujian terhadap *nonword error detection* dan *isolated-word error correction* menggunakan data dari ICL-*corpus*, akan dilakukan analisis sesuai dengan masalah yang telah dirumuskan dalam bagian rumusan masalah.

1.6 Sistematika Penulisan

Guna memudahkan dalam mendapatkan gambaran yang lengkap dan jelas mengenai penelitian yang akan dilakukan, penulis membagi laporan ini menjadi 5 (lima) bab yaitu Bab I Pendahuluan, Bab II Tinjauan Pustaka, Bab III Analisis dan

Perancangan Sistem, Bab IV Implementasi dan Analisis Sistem, dan Bab V Kesimpulan dan Saran.

Bab 1 menguraikan hal-hal seperti latar belakang masalah, perumusan masalah, batasan masalah, tujuan penelitian, metode/pendekatan yang digunakan serta sistematika penulisan laporan tugas akhir.

Bab 2 berisi tentang tinjauan pustaka serta landasan teori yang diperlukan untuk memecahkan masalah dalam riset yang dilakukan.

Bab 3 berisi tentang kebutuhan sistem yang dibangun dalam penelitian, *flowchart* dan arsitektur sistem, perancangan evaluasi terhadap sistem, dan contoh perhitungan manual.

Bab 4 berisi tentang hasil penelitian/implementasi serta pembahasan/analisis dari penelitian yang telah dilakukan dan dijelaskan secara terpadu.

Bab 5 berisi kesimpulan dari sistem yang telah dibuat dan saran yang akan berguna untuk pengembangan sistem selanjutnya. Dengan adanya saran, diharapkan riset yang dilakukan selanjutnya akan menghasilkan hasil yang lebih baik.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil implementasi dan analisis sistem, maka diperoleh kesimpulan sebagai berikut:

1. Dalam analisis *nonword error detection*, didapat nilai *precision*, *recall*, dan *f-measure* sebesar 0.97716, 0.77621, dan 0.86517.
2. Secara khusus, performa koefisien Dice tidak baik. Pernyataan tersebut dibuktikan dengan nilai *f-measure* koefisien Dice yang lebih rendah jika dibandingkan dengan metode lain yang juga menggunakan Damerau-Levenshtein dan frekuensi kata. Perbandingan nilai *f-measure* metode yang hanya menggunakan koefisien Dice dengan yang menggunakan Damerau-Levenshtein dan frekuensi kata dapat dilihat datanya pada Tabel 4.38.
3. Performa metode yang menggunakan k-gram kombinasi lebih baik daripada metode yang menggunakan k-gram biasa. Pernyataan tersebut dibuktikan dengan nilai *f-measure* yang lebih tinggi jika menggunakan k-gram kombinasi. Perbandingan nilai *f-measure* secara lengkap dapat dilihat pada Tabel 4.38.
4. Dalam hasil evaluasi *isolated-word error correction*, metode yang paling baik performanya adalah metode bigram kombinasi yang menggunakan koefisien Dice, Damerau-Levenshtein, dan frekuensi kemunculan *term* pada *threshold* koefisien Dice sebesar 0.2 dengan nilai *precision*, *recall*, dan *f-measure* sebesar 0.56345, 0.44758, dan 0.49888. Hasil evaluasi tersebut membuktikan bahwa penggunaan algoritma Damerau-Levenshtein dan *word frequency* dapat meningkatkan performa sistem untuk melakukan *isolated-word error correction*.
5. Teknik *nonword error detection* dan *isolated-word error correction* memiliki ketergantungan yang tinggi terhadap kamus dan daftar kata. Sistem yang

dibangun memerlukan kamus bahasa Indonesia, kamus bahasa Inggris, kamus akronim dan singkatan, daftar kata nonformal, dan daftar *proper nouns*. Tidak sedikitnya jumlah kamus dan daftar kata yang digunakan menunjukkan ketergantungan sistem yang tinggi terhadap kamus dan daftar kata.

5.2 Saran

Saran untuk pengembangan dan perbaikan sistem adalah sebagai berikut:

1. Pada teks dokumen uji, terdapat *term* salah yang menggunakan imbuhan bahasa Indonesia. Sedangkan dalam kamus bahasa Indonesia yang digunakan, bentuk kata yang memiliki imbuhan tidaklah lengkap. Oleh karena itu, pada tahap *isolated-word error correction* perlu digunakan analisis morfologi untuk mengatasi jenis kesalahan tersebut.
2. Pada teks dokumen uji, terdapat kesalahan berupa tidak digunakannya spasi sebagai pemisah antara dua kata, contohnya seperti kesana dan waktunonton. Oleh karena itu, diperlukan penggunaan metode yang dapat menyisipkan spasi di antara dua kata yang tidak menggunakan spasi dengan memperhatikan konteks kedua kata tersebut supaya juga dapat mengetahui apakah penggunaan kata depan pada suatu kata adalah benar atau salah penulisannya.
3. Berdasarkan pengamatan terhadap teks dokumen uji, terdapat kesalahan berupa penggabungan unsur bahasa asing dengan unsur bahasa Indonesia, sehingga diperlukan penggunaan metode untuk menyisipkan tanda hubung di antara unsur bahasa asing dan unsur bahasa Indonesia.
4. Dikarenakan terdapat kesalahan penggunaan tanda baca, diperlukan penggunaan metode untuk memeriksa dan mengoreksi kesalahan dalam hal penggunaan tanda baca seperti tanda titik,tanda petik, dan tanda hubung yang melekat pada suatu *term.*
5. Penggunaan metode yang lebih baik untuk mengatasi kata ulang yang disingkat karena beragamnya bentuk singkatan kata ulang pada teks dokumen uji.

6. Analisis mengenai kapan harus menggunakan koefisien Dice dan Damerau-Levenshtein.

DAFTAR PUSTAKA

Ardriyati, W. (2012). A Lexical Study On Non Standard Expressions In Students' Facebook. *Dinamika Bahasa & Budaya Volume 7, No.2* .

Atmajaya, G. E. (2008). *Pembuatan Spelling Checker untuk Bahasa Indonesia dengan Java 2 Standard Edition.* Depok.

Bentley, J. L., & Sedgewick, R. (1997). Fast Algorithms for Sorting and Searching Strings.

Boriah, S., Chandola, V., & Kumar, V. (2008). *Similarity Measures for Categorical Data: A Comparative Evaluation.* Minnesota.

Christen, P. (2012). *Data Matching : Concepts and Techniques for Record Linkage, Entity Resolution and Duplicate Detection.* Heidelberg: Springer.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology , 26*, 297-302.

Dwitiyastuti, R. N., Muttaqin, A., & Aswin, M. (2013). Pengoreksi Kesalahan Ejaan Bahasa Indonesia Menggunakan Metode Levenshtein Distance. *Jurnal Mahasiswa TEUB , 1*.

Hassal, D. T. (2012). *Learning to Read Colloquial Indonesian.*

Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys (CSUR) , 24* (4), 377-439.

Lee, J. S. (2009). *Automatic Correction of Grammatical Errors in Non-native English Text.* Cambridge: Massachusetts Institute of Technology.

Luqman. (2009). Program Aplikasi Pengoreksian Ejaan Bahasa Indonesia. *PETIR , 2*, 13-19.

May. (2011, February 14). *Reading*. Dipetik November 06, 2014, dari 3rd year

project: http://mayana2011.blogspot.com/2011/02/reading_14.html

Min, K., Wilson, W. H., & Moon, Y.-J. (2000). Typhographical And Orthographical Spelling Error Correction. *International Conference on Language Resources and Evaluation.* Athens.

Pfeifer, U., Poersch, T., & Fuhr, N. (1994). Searching Proper Names in Databases.

Radovanovic, M., & Ivanovic, M. (2008). Text Mining : Approaches and Applications. *Novi Sad Journal Of Mathematics* , 227-234.

Sarpong, K. A.-M., Davis, J. G., & Panford, J. K. (2013). A Conceptual Framework for Data Cleansing – A Novel Approach to Support the Cleansing Process. *International Journal of Computer Applications* .

Sutisna, U. (2009). *Koreksi Ejaan Query Bahasa Indonesia Menggunakan Algoritme Damerau Levenshtein.*