PENGOREKSIAN EJAAN KATA MENGGUNAKAN METODE N-GRAM (STUDI KASUS DOKUMEN TEKS BERBAHASA INDONESIA)

Wedha Satya Wardhana[1], Tjokorda Agung Budi Wirayuda[2], Shaufiah[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

Abstrak

Kemudahan berkomunikasi saat ini menjadikan terjadinya pertukaran informasi yang sangat cepat, berbagai media informasi yang ada mulai dari televisi, Koran dan internet masih memanfaatkan tulisan sebagai salah satu cara menyampaikan informasi. Namun kadangkala informasi tersampaikan tidak sempurna karena adanya kendala diantaranya karena kesalahan penulisan secara tidak disengaja, oleh karena itu diperlukan adanya solusi pengecekan kata.

Proses pengoreksian kata terdiri atas 2 langkah, yaitu pengecekan kata yang salah dimana setiap kata yang diketik akan dicari apakah kata tersebut termasuk kata baku kemudian menandai kata yang salah eja. Langkah berikutnya yaitu mencari kata yang tepat untuk menggantikan kata yang salah tersebut.

Dalam tugas akhir ini sistem yang dibangun mengimplementasikan proses pengoreksian ejaan kata untuk bahasa Indonesia menggunakan bi-gram dan trigram kemudian menguji hasil koreksi dokumen. Parameter hasil pengujian meliputi waktu eksekusi dan jumlah kata yang dapat dikoreksi secara benar dengan mengunakan sistem ini. Pengujian menghasilkan akurasi mendekati sempurna untuk pengujian menggunakan user interaksi, dan untuk pengujian secara otomatis system menghasilkan akurasi dibawah 45 persen, namun dengan penggunaan threshold pada sistem, dapat menaikkan akurasi dengan rata-rata hasil diatas 60 persen. Untuk metode n-gram yang dilakukan, secara umum akurasi bigram sedikit lebih baik daripada trigram, adapun waktu respon sistem untuk memproses setiap kata salah berkisar antara 1~3 detik, sehingga dapat dikatakan respon sistem cukup tanggap dan dapat ditolerir oleh user pengguna.

Kata Kunci : spell correction, spell checker, spell suggestion, bigram, trigram.

Abstract

The easiness of communication makes the exchange of information very quickly, many information media ranging from television, newspapers and the Internet are still use writing as a way to convey information. Yet, sometimes the information conveyed through is still imperfect because of some constraints such as a typing error, therefore it need a solution to check for word error.

The process of word correction consists of two steps. First, check the error word in which every word you typed will be checked, if the word is considered as not a part of real word it then mark the word as misspelled words. The next step is to find the correct words to replace the error word.

In this final assignment, a system was built to implement words corrector for the Indonesian language documents by using bi-grams and trigrams method, then the system would test the results of the corrected documents. Parameter of test results including execution time and number of words that can be corrected properly by using this system. Test resulting in near perfect accuracy if used with user interaction, a little below 45 percent in automated, but it can greatly improved accuracy resulting by an average of above 60 percent if used in automated with threshold enabled. As for n-gram methods used, bigram is resulting slightly higher accuracy than trigram, at the expense of longer response time than trigram. And for the average system response time, it take about 1 ~ 3 seconds to process every error word, so it quite responsive and can still be tolerated by the normal user.

Keywords : spell correction, spell checker, spell suggestion, bigram, trigram.

1. Pendahuluan

1.1 Latar Belakang

Bahasa adalah salah satu komponen penting dalam kehidupan manusia, dimana dengan bahasa, seseorang dapat menyalurkan pemikirannya kepada orang lain[15]. Penggunaan bahasa sesuai dengan kaidah yang telah diterapkan merupakan suatu hal yang sangat diharapkan. Namun ada kalanya dalam menggunakan bahasa dalam bentuk tulisan, manusia tanpa disadarinya membuat kesalahan dalam pengejaan sehingga mengaburkan pengertian dari kalimat yang dituliskan.

Penelitian mengenai pengecekan ejaan pertama kali dilakukan oleh sekumpulan ahli bahasa pada tahun 1970-an. Sejak itu, mulai berkembang penelitian-penelitian mengenai pengecekan ejaan di seluruh dunia sampai sekarang[12]. Pengoreksian ejaan kata adalah proses tiga langkah. Pertama melakukan pengecekan kata (*spell checking*) terhadap kata dalam kamus. Kedua, menyeleksi kata saran yang potensial (*spell suggestion*) sebagai kata yang benar. Ketiga, mengurutkan kata saran *(sorting)* dimana diharapkan urutan teratas adalah kata yang diinginkan[6].

Beberapa metode telah dikembangkan untuk membantu proses pengoreksian ejaan kata, metode yang paling umum digunakan adalah *Hamming Distance* dan *Levenshtein Distance*[2]. *Hamming Distance* adalah metode pengoreksian dengan cara membandingkan dua string yang panjangnya sama, dan menghitung jarak kesalahan minimal dengan melakukan subtitusi sampai diperoleh string yang tepat. *Hamming Distance* umumya digunakan untuk proses searching serta error detecting dan error corrrection code pada bidang telekomunikasi. Namun kekurangan metode ini adalah kata yang diperbandingkan harus memilki panjang yang sama[5]. Levenshtein Distance adalah metode pengembangan Hamming Distance, dimana menghitung jarak kesalahan minimal dengan cara membandingkan dua string dengan melakukan insert, delete, dan subtitusi sehingga mampu menangani string yang memiliki panjang berbeda. Namun kelemahan metode ini terletak pada proses penghitungan distance antara dua buah kata, dimana setiap huruf yang ada pada kedua kata akan dibandingkan dan dilakukan proses subtitusi, insertion, delete hingga kedua kata tersebut sama lalu didapatkan distance (jumlah perbedaan) antara kedua kata tersebut[7]. Proses ini terus dilakukan untuk seluruh kata yang akan dibandingkan, hingga didapat hasil akhir berupa kandidat kata dengan distance terkecil dari seluruh kata yang ada, kandidat kata tersebut akan menjadi kata yang digantikan. Sehingga jika kata yang dibandingkan jumlahnya sangat banyak, proses ini akan banyak memakan waktu[11].

N-gram adalah salah satu alternatif untuk menyelesaikan masalah pengoreksiaan ejaan kata. N-gram adalah segmen teks yang terdiri dari n-karakter, termasuk pemisah

antar kata (biasanya berupa spasi). Pendekatan N-Gram yaitu menggunakan *dictionary look-up* dengan cara membandingkan n-gram dari kata yang salah dengan n-gram kata yang ada pada database yang telah diinisialisasi sebelumnya. Dikarenakan proses pembuatan n-gram kata hanya dilakukan satu kali sebelum aplikasi dijalankan, dan setelahnya n-gram kamus tersebut dapat digunakan berulangkali, sehingga diharapkan waktu pemrosesan akan singkat. Keuntungan lain dari pendekatan ini adalah hubungan antar-huruf dapat dijaga, yang biasanya tidak tercakup bila dilakukan pendekatan per kata. [14]

Tugas akhir ini akan menghasilkan sistem pengoreksian kata dengam menggunakan bagian dari N-gram, yaitu metode *Bigram dan Trigram*, kemudian menganalisa hasil performansi dari sistem. Sistem yang dibangun diharapkan mempunyai tingkat akurasi yang tinggi dan waktu respon sistem yang masih dapat diterima oleh user pengguna.

1.2 Perumusan Masalah

Seperti yang telah dipaparkan sebelumnya, permasalahan yang dihadapi adalah proses pengoreksian ejaan kata dengan menggunakan penerapan N-gram, namun proses ini harus mampu menyelesaikan masalah yang ada yaitu:

1. Bagaimana implementasi pengoreksian meliputi proses *spell checking, dan spell suggestion* sehingga dapat menemukan kata yang salah ejaan, kemudian memperbaikinya.
2. Bagaimana penerapan algoritma *N-gram* dapat memberikan usulan ejaan kata yang seharusnya benar.

1.3 Batasan masalah

Agar masalah tidak terlalu luas, maka batasan masalah dalam tugas akhir ini antara lain:

1. Pengimplementasian Metode N-gram pada tugas akhir ini dibatasi pada Bigram(n=2) dan Trigram *(n=3)* pada level subkata.
2. Teks sumber yang akan diuji berbahasa Indonesia, dan di utamakan berasal dari berita dan karya tulis untuk kepastian struktur bahasa yang baik.
3. Daftar usulan kata diambil berdasarkan dari daftar kata yang terdapat pada *Kamus Besar Bahasa Indonesia,* Jakarta:Departemen Pendidikan Nasional,2008.
4. Pengoreksian dilakukan dengan membandingkan pada kata yang terdapat pada kamus (*non word error*).
5. Pengoreksian tidak memperhatikan grammar, maupun stuktur kata.
6. Koreksi untuk huruf, tidak untuk angka maupun tanda baca.

1.4 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Menghasilkan sistem pengoreksian ejaan kata yang mengimplementasikan metode N-gram yang dapat melakukan pengoreksian ejaan kata terhadap kumpulan dokumen berbahasa Indonesia, khususnya artikel berita.
2. Menganalisis pengaruh penggunaan Threshold terhadap hasil akurasi pengoreksian yang dilakukan secara otomatis oleh sistem.
3. Menganalisis performansi berupa jumlah kata yang dapat dikoreksi dengan tepat (akurasi) dan waktu pemberian kandidat kata dari implementasi sistem pengoreksian.

Hipotesa Awal:

Berdasarkan atas paper yang ditulis oleh Manirul (2006) berjudul *Analysis of N-Gram Based Text Categorization for Bangla in a Newspaper Corpus* diperoleh kesimpulan bahwa trigram(n=3) menghasilkan akurasi yang paling tinggi. Meskipun permasalahan yang dihadapi antara bahasa Bangla dan Bahasa Indonesia berbeda, namun dapat diambil dugaan awal, bahwa Trigram tetap akan memberikan hasil terbaik, dan jumlah kata turut mempengaruhi hasil pengujian akurasi.

1.5 Metodologi Penyelesaian Masalah

Metodologi penyelesaian masalah yang dilakukan antara lain:

1. Studi Literatur.
   Mempelajari dan mengumpulkan literature yang relevan dengan permasalahan dan tujuan penelitian untuk mendapatkan dekripsi yang jelas dan dasar teori mengenai metode pengoreksian ejaan kata serta dan metode N-gram.

2. Analisis dan Perancangan
   Menganalisa requirement berdasarkan permasalahan dan kemungkinan modifikasi atau penambahan terhadap algoritma pengoreksian ejaan kata berbasis N-gram yang akan diimplementasikan. Pada tahap ini pula perancangan perangkat lunak pengoreksiaan ejaan kata menggunakan implementasi N-gram dengan menggambarkan semua aspek perangkat lunak meliputi desain database, dan desain struktur data.

3. Implementasi Coding
   Pada proses ini, design yang telah dihasilkan akan ditransformasikan kedalam bentuk yang dimengerti oleh mesin. Aplikasi antarmuka dibangun dengan menggunakan compiler Visual C# dan database dibuat menggunakan MYSQL.

4. Testing dan Evaluasi
   Di bagian ini dimana design dibentuk melalui coding. Yang kemudian akan di tes secara terpisah. Bila masing-masing coding telah di tes maka akan di lakukan pengetes-an secara menyeluruh. Bila mengalami error, akan dilakukan pengecekan dan perbaikan. Adapun pengujian metode dilakukan dengan membandingkan dokumen yang terdiri dari paragraph yang memiliki kata yang salah (d1`) dan telah dikoreksi menggunakan implementasi metode N-gram terhadap dokumen yang telah diperiksa ejaannya secara manual (d1).

5. Penyusunan Laporan dan Kesimpulan Akhir.
   Mencatat setiap proses yang dilakukan, kemudian menyusunnya kedalam laporan dan mengambil kesimpulan akhir dari keseluruhan penelitian tugas akhir Pengoreksiaan Ejaan Kata menggunakan implementasi metode N-gram.

1.6 Sistematika Penulisan

Tugas akhir ini disusun berdasarkan sistematika sebagai berikut :

Bab I **Pendahuluan**
Bab ini akan membahas kerangka penelitian atau percobaan dalam tugas akhir, meliputi latar belakang masalah, perumusan masalah, tujuan, batasan masalah, metode penyelesaian masalah, dan sistematika penulisan.

Bab II **Dasar Teori**
Bab ini memuat berbagai dasar teori yang mendukung dan mendasari penulisan tugas akhir ini, yaitu mengenai konsep dari pengoreksian ejaan kata meliputi *spell checker* dan spell *suggestion*. Serta konsep dari *n-gram*.

Bab III **Analisis Perancangan dan Implementasi**
Berisi analisis sistem pengoreksian menggunakan n-gram yang akan dibuat mencakup analisis kebutuhan sistem, *flow-chart* perancangan proses dan implementasinya sehingga dapat memahami sistem secara jelas.

Bab IV **Pengujian**
Berisi tentang hasil pengujian sistem pengoreksian menggunakan n-gram yang telah dibuat.

Bab V **Kesimpulan dan Saran**
Berisi tentang kesimpulan dari keseluruhan aplikasi yang dibuat serta saran untuk pengembangan selanjutnya.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Berdasarkan hasil analisis terhadap hasil pengujian dapat disimpulkan sebagai berikut:

1. Metode bigram dan trigram dapat diimplementasikan untuk melakukan pengoreksian ejaan kata dalam teks berbahasa Indonesia.
2. Pengoreksian ejaan dengan metode n-gram bisa memberikan hasil yang baik saat digunakan dengan bantuan dari user untuk memilih kata saran ditunjang dengan waktu eksekusi yang masih cukup cepat, namun masih kurang sempurna pengoreksiannya jika selurunnya dilakukan otomatis oleh sistem.
3. Penggunaan Threshold pada pengoreksian otomatis meningkatkan persentase ketepatan kata secara signifikan, dengan waktu eksekusi yang tak berbeda dengan pengoreksian otomatis tanpa threshold.
4. Performansi sistem pengoreksian menggunakan implementasi metode bigram lebih tinggi dibandingkan menggunakan trigram, namun waktu eksekusi sistem pengoreksian trigram lebih singkat dibandingkan dengan bigram.
5. Jenis kesalahan kata mempengaruhi performansi, error kata yang disebabkan oleh insertion dan deletion dapat dikoreksi dengan baik oleh algoritma n-gram, namun algoritma n-gram kurang baik mengoreksi error kata yang disebabkan oleh substitusi dan transposisi.

5.2 Saran

Berdasarkan hasil saran dan kesimpulkan, terdapat beberapa saran untuk perbaikan penelitian pengoreksian kata:

1. Untuk meningkatkan persentase kata yang benar, bisa dilakukan pengoreksian terhadap struktur kalimatnya, diantara dengan menggunakan n-gram pada level sub kalimat.

Daftar Pustaka

[1] Amber Wilcox-O'Hearn, Graeme Hirst, and Alexander Budanitsky. *Real-word spelling correction with trigrams*: A reconsideration of the Mays, Damerau, and Mercer model. Toronto University, Canada. 2008

[2] Douglass, Robert. *Add spelling suggestions to the Search Result Page*. http://drupal.org/node/247482.

[3] Gilleland, Michael. *Levensthein Distance*. http://www.merriampark.com/ld.htm

[4] Hanafi, Ahmad. *Pengenalan Bahasa Suku Bangsa Indonesia Berbasis Teks Menggunakan Metode N-gram*. ITTelkom, Bandung, 2009

[5] Hasan, Ahmad, Sarah Neuman, Hany Hassan. *Language Independent Text Correction Using Finite State Automata*. IBM Cairo Technology Development Center.

[6] Karen, Kukich. 1992b. *Techniques for automatically correcting words in text*. A CM Computing Survey

[7] Levenshtein, V.I. 1966. *How Levenshtein Works*. http://www.levenshtein.net/ index.html

[8] Miller, Robert B. *Response time in man-computer conversational transactions*. International Business Machines coporation, Poughkeepsie, New York.1968

[9] Priyadi.http://priyadi.net/archives/2008/02/19/kamus-besar-bahasa-indonesia-kbbi-versi-internet/

[10] R. C. Angell, G. E. Freund and P. Willett. *Automatic spelling correction using trigram similarity measur*e. Information Processing and Management. 1983.

[11] _____. *How Spellchekers Work*. http:// pcplus.techradar.com/node/3062

[12] _____.*Spellchecking by Computer*. http://www.dcs.bbk.ac.uk/~roger/spellchecking.html

[13] _____.*What is Spell Checker?* http://www.compassrose.com/publishing/about-spell-checker.html

[14] _____.http://oguds.wordpress.com/2007/11/17/pencarian-semantik-web-menggunakan-swoogle/

[15] _____.http://digilib.petra.ac.id/jiunkpe/s1/info/2004/jiunkpe-ns-s1-2004-26499067-4398-bahasa-chapter1.pdf