Jurnal Ilmiah Ilmu Komputer, Vol 15 No. 2, Desember 2010 : 25 - 29

Koreksi Ejaan Query Bahasa Indonesia Menggunakan Algoritme Damerau Levenshtein

Utis Sutisna, Julio Adisantoso

Departemen Ilmu Komputer, Institut Pertanian Bogor, Jl. Meranti Wing 20 Lv.V, Bogor, Jawa Barat, 16680

Abstract---Query spelling on search engine is important to improve the quality information searching result. When user types query for search engine input, sometime spelling mistakes occurred due to position of keyboard and finger movement while typing. As an effect, searching result is incorrect and when user misspells the query, information obtained will not succeed. Therefore, search engine requires an application of spelling corrections. This research proposes correction process for query spelling by giving words suggestions which are obtained by calculating edit distance for each corrected word towards every word in dictionary. The concept for calculating edit distance uses Damerau Levenshtein algorithm which consists of 4 operations: (1) insertion, (2) substitution, (3) deletion, and (4) transposition. This research shows that implementation of Damerau Levenshtein algorithm is able to increase recall-precision value in information retrieval system. It is shown by the increasing of average recall-precision value at 44,82% after correction.

Keywords : Damerau Levenshtein, Algoritme Damerau Levenshtein Metric, edit distance

PENDAHULUAN

Kata kunci atau yang biasa disebut dengan *query* pada pencarian informasi dari sebuah *search engine* digunakan sebagai kriteria pencarian yang tepat dan sesuai dengan kebutuhan. Ejaan kata kunci yang benar menjadi penting untuk meningkatkan hasil pencarian informasi. Ketika pengguna menulis *query* sebagai masukan pada sistem pencari, muncul kesalahan ejaan disebabkan posisi tombol papan ketik dan pergerakan jari sehingga hasil pencarian bersifat salah. Oleh karena itu, diperlukan suatu aplikasi yang dapat mengoreksi kesalahan ejaan. Pengoreksian ejaan ini dapat dilakukan dengan memberikan ejaan kata yang benar yaitu dengan memberikan usulan ejaan kata yang mirip berdasarkan kamus.

Penelitian tentang pengoreksian ejaan bahasa Indonesia juga pernah dilakukan oleh Primasari (1997), melakukan penelitian tentang pencarian dan temu kembali nama berdasarkan kesamaan fonetik. Arumsari (1998) dengan menggunakan metode jarak *edit*. Wahyudin (1999) dengan menggunakan algoritme trigram untuk mendapatkan kata-kata perkiraan dari kata yang dinyatakan salah eja. Arumsari (1998) menentukan jarak *edit* diantara dua *string* dari operasi yang dilakukan yaitu: (1) operasi penyisipan (*insertion*), (2) operasi penghapusan (*deletion*), dan (3) operasi penggantian (*subtitution*) sebuah huruf. Pada penelitian ini, pengoreksian ejaan akan dilakukan dengan algoritme Damerau Levenshtein dengan metode jarak *edit*. Operasi yang dilakukan tidak hanya tiga operasi seperti yang dilakukan dalam penelitian Arumsari (1998). Tetapi, juga diperhatikan operasi penukaran (*transposition*) posisi sebuah huruf yang berdekatan. Sehingga perolehan kata ejaan yang benar lebih optimal.

METODOLOGI

Menurut Damerau dalam Wahyudin (1999) menyimpulkan 80% kesalahan ejaan dapat disebabkan karena empat hal, yaitu:

1. Penggantian satu huruf
2. Penyisipan satu huruf
3. Penghilangan satu huruf
4. Penukaran dua huruf berdekatan.

Kesalahan ejaan juga dapat disebabkan oleh beberapa hal, diantaranya:

1. Ketidaktahuan penulisan. Kesalahan ini biasanya konsisten dan kemungkinan berhubungan dengan bunyi kata dan penulisan yang seharusnya.
2. Kesalahan dalam pengetikan yang lebih tidak konsisten tapi mungkin berhubungan erat dengan posisi tombol papan ketik dan pergerakan jari.
3. Kesalahan transmisi dan penyimpanan yang berhubungan dengan pengkodean pada jalur mekanisme transmisi data.

Kesalahan ejaan dapat dikoreksi menggunakan dua strategi dasar yang berbeda, yaitu mutlak dan relatif (Pullock & Zamora 1984, dalam Wahyudin 1999). Secara mutlak, pengoreksian dilakukan dengan membuat suatu tabel variasi ejaan yang salah dengan ejaan yang benarnya. Namun demikian, secara relatif ejaan yang benar dipilih dari kamus yaitu dengan mencari kata dalam kamus yang paling mirip dengan kata yang salah ejaannya.

Damerau Levenshtein Metric adalah sebuah fungsi pada *finite* string dari sebuah alphabet ke integer. Sebuah matriks jarak yang diberikan *strings s1, s2, s3* yang memenuhi kondisi (Bard 2006):

- *Non-negativity*: $d(s_1, s_2) \geq 0$
- *Non-degeneracy*: $d(s_1, s_2) = 0$ jika dan hanya jika $s_1 = s_2$
- *Symmetry*: $d(s_1, s_2) = d(s_2, s_1)$
- *Triangle Inequality:* $d(s_1, s_2) + d(s_2, s_3) \geq d(s_1, s_3)$

Jarak $d(s_1, s_2)$ didefinisikan sebagai sebuah kombinasi operasi penjumlahan dari penambahan sebuah huruf, penghilangan sebuah huruf, penggantian sebuah huruf atau penukaran sebuah huruf dari huruf lainnya dalam satu lokasi.