Perbandingan Penggunaan Kamus Terdistribusi, Partition Around Medoids (PAM) dan Struktur Data Trie dalam Perbaikan Ejaan Otomatis Pada Teks Formal Bahasa Indonesia = A Comparison of Distributed, PAM, and Trie Data Structure Dictionaries in Automatic Spelling Correction for Indonesian Formal Text

Mukhlizar Nirwan Samsuri, author

Abstrak

Kesalahan ejaan dapat dibagi menjadi dua jenis, non-word errors dan real-word errors. Non-word errors adalah kesalahan eja yang tidak terdapat dalam kamus, sedangkan real-word errors adalah kata yang terdapat pada kamus tetapi berada pada tempat yang tidak tepat pada kalimat. penelitian ini berfokus pada koreksi ejaan untuk non-word errors pada teks formal Bahasa Indonesia. Tujuan dari penelitian ini adalah untuk membandingkan efektivitas tiga jenis struktur kamus untuk koreksi ejaan, antara lain kamus terdistribusi, kamus PAM (Partition Around Medoids), dan kamus menggunakan struktur data trie. Ketiga jenis kamus juga akan dibandingkan dengan kamus sederhana yang dijadikan sebagai baseline. Tahap pengurutan kandidat (ranking correction candidates) dilakukan dengan menggunakan dua variasi dari edit distance, yaitu Levenshtein dan Damerau-Levenshtein dan n-gram. Guna mendukung penelitian ini, dibangun dataset gold standard dari 200 kalimat yang terdiri dari 4.323 token dengan 288 di antaranya adalah non-word errors. Berdasarkan kombinasi tipe kamus dan edit distance, didapatkan hasil bahwa struktur data trie dengan Damerau-Levenshtein distance memperoleh accuracy terbaik untuk menghasilkan kandidat koreksi, yaitu 95,89% dalam 45,31 detik. Selanjutnya, kombinasi struktur data trie dengan Damerau-Levenshtein distance juga mendapatkan accuracy terbaik dalam memilih kandidat terbaik, yaitu 73,15%.
.....Spelling errors can be divided into two groups: non-word and real-word. A non-word error is a spelling error that does not exist in the dictionary, while a real-word error is a real word but not on the right place. In this work, we address the non-word errors in spelling correction for Indonesian formal text. The objective of our work is to compare the effectiveness of three kinds of dictionary structure for spelling correction, distributed dictionary, PAM (Partition Around Medoids) dictionary, and dictionary using trie data structure, with the baseline of a simple flat dictionary. We conducted experiments with two variations of edit distances, i.e. Levenshtein and Damerau-Levenshtein, and utilized n-grams for ranking correction candidates. We also build a gold standard of 200 sentences that consists of 4,323 tokens with 288 of them are non-word errors. Among the various combinations of dictionary type and edit distance, the trie data structure with Damerau-Levenshtein distance gets the best accuracy to produce candidate correction, i.e. 95.89% in 45.31 seconds. Furthermore, the combination of trie data structure with Damerau-Levenshtein distance also gets the best accuracy in choosing the best candidate, i.e. 73.15%.