PENGGUNAAN SPELLING CORRECTION DENGAN METODE PETER NORVIG DAN N-GRAM

Ricky Martin [1], Dali Santun Naga [2], Viny Christanti Mawardi [3]

Teknik Informatika Universitas Tarumanagara
Jl.Letjen S Parman No. 1, Grogol Petamburan, Jakarta Barat 11440 Indonesia
ricky.535170048@stu.untar.ac.id [1], dalinaga@gmail.com [2], vinym@fti.untar.ac.id [3]

ABSTRACT

Typing errors in a document are human errors that are difficult to avoid as a result of which the message to be conveyed is not optimal. Using the Spelling Corrector feature is one way to check typing errors. The Spelling Corrector feature is able to provide only one word suggestion and correct it immediately. The combination of Peter Norvig and Ngram's methods was able to come up with one word suggestion and correct it right away. Both methods of looking for word suggestions use the probability value of the words that appear most frequently in the dictionary. The difference between the two methods is that Peter Norvig uses an algorithm that combines the delete, insert, replace, and transpose processes for the wrong word. Meanwhile, Ngram uses an algorithm that pays attention to the words before and after them based on the sentences in the dictionary. The dictionary used is a dictionary of word lists from KBBI and also a dictionary of sentences taken from several articles on the internet. This combination of methods was tested using 55 documents containing sentences that had one misspelled word. The test results show that the combination of the two methods provides an accuracy rate of 73,684% and a success rate of 37,037% for the total accuracy of this application is 69.09%. These two methods can be used to correct typing errors, although they cannot correct words with an error rate of two or more letters. This is because Peter Norvig was unable to correct a word with a two-letter error rate and needed a good corpus.

Key words

Kombinasi, N-Gram, Peter Norvig, Spelling Corrector

1. Pendahuluan

1.2. Latar Belakang

Bahasa juga dapat dijadikan acuan dalam penulisan dokumen, komunikasi dan pencarian informasi. Apabila penulisan yang terdapat pada dokumen adanya tipografi, maka penulisan dapat membuat arti dari kata yang disampaikan menjadi keliru atau memiliki arti lain.[1]

Pada aplikasi *Spelling Correction* atau *Spelling Checker* ini mempunyai fungsi yaitu mencari kata yang mengalami kesalahan ejaan berdasarkan kata korpus yang digunakan terhadap aplikasi serta berfungsi untuk memberikan saran kata yang dilakukan dengan algoritma yang juga digunakan oleh aplikasi. Sementara pengoreksian kata yang terdapat kesalahan ejaan (*Spelling Correction*) adalah sistem yang dibuat untuk mendeteksi kesalahan ejaan dan memperbaikinya.

Pada perancangan aplikasi *Spelling Checker* atau *Spelling Correction* akan menggunakan metode *Peter Norvig*. Metode *Peter Norvig* merupakan metode yang dibuat oleh perusahaan search engine dan algoritma unik yang dikombinasikan dengan proses menghapus, menambah, mengubah, dan mengganti huruf pada kata yang salah. Kata tersebut harus dicek kembali pada kamus. Sedangkan sistem *Spelling Correction* akan dibuat menggunakan metode N-gram dengan perhitungan pemenggalan kata 2-gram (Bigram) menggunakan *Language Model. Spelling Correction* maupun *Spelling Checker* mempunyai 2 macam jenis kesalahan ejaan yaitu kesalahan penulisan kata sah (real word error) dan kesalahan kata tidak sah (non word error). Kesalahan kata sah (real word error) merupakan kesalahan pada kata sehingga memiliki makna lain, contohnya penulisan "kasur" menjadi "kapur". Sedangkan kesalahan kata tidak sah (non word error) merupakan kesalahan pada kata sehingga menjadi tidak bermakna, contohnya penulisan "kasur" menjadi "ksur". [2]

1.3. Rumusan Rancangan

Rancangan aplikasi yang digunakan untuk aplikasi *Spelling Correction* menggunakan metode Peter Norvig dan N-gram ini berbasis website aplikasi sehingga dapat mudah diakses dan digunakan. Tampilan dari aplikasi dirancang dengan

menggunakan bahasa pemrograman *PHP*. dan untuk perhitungan *Spelling Correction* menggunakan bahasa pemrograman *Python*.

1.3. Tujuan Rancangan

Tujuan dari perancangan aplikasi *Spelling Correction* sebagai berikut:

1. Mempermudah pengguna untuk mengoreksi dokumen berupa .txt dan mengubah kata-kata yang mengandung tipografi.
2. Membuat aplikasi *Spelling Correction* berbasis web untuk mengoreksi dan mengubah kata-kata di dalam dokumen ketika terdapat ejaan yang salah dalam Bahasa Indonesia.
3. Membuat aplikasi *Spelling Correction* untuk mengoreksi dan mengubah kata-kata di dalam dokumen dengan benar.

2. Landasan Teori

2.1 Spelling Correction dan Spelling Checker

Spelling Correction merupakan proses mendeteksi, mengoreksi, dan memberikan saran kata untuk kata-kata yang mengalami kesalahan pada ejaan di dalam suatu teks. Sedangkan *Spelling Checker* merupakan sistem yang akan melakukan proses pencarian kata-kata yang salah ejaan berdasarkan penggunaan data korpus yang dirancang.

2.2. Dictionary Lookup

Metode *dictionary lookup* merupakan metode yang digunakan dalam penentuan non-word error. Proses yang dilakukan adalah pengecekan kata yang terdaftar dalam kamus atau tidak, jika tidak terdapat dalam kamus maka dianggap sebagai non-word. Cara ini adalah cara yang efektif untuk penentuan kata yang termasuk adanya kesalahan penulisan atau tidak, namun, jumlah kata yang banyak dalam kamus dapat mengakibatkan proses pengecekan kalimat menjadi sangat lama. [3] Permasalahan lain dalam menggunakan kamus yaitu terdapat kata yang sudah diberikan imbuhan pada kamus, serta terdapatnya kata asing, turunan kata dan kata baru yang tidak bisa diprediksi kemunculannya pada dokumen, selain itu terdapat beberapa bidang seperti kesehatan, ekonomi, biologi mempunyai istilah khusus yang terkadang tidak terdapat didalam kamus umum.

2.3. Metode Peter Norvig

Peter Norvig merupakan sistem dengan mengubah jarak kata salah ejaan atau mengubah kata yang salah menjadi dua kata dan sejumlah suntingan yang membutuhkan kata untuk mengubah satu ke yang lain. Pendekatan kata pada metode *Peter Norvig* dapat menghasilkan semua kemungkinan kata dengan semua operasi edit-distance yaitu operasi penambahan (insert), operasi penggantian (replace), operasi penukaran (transpose), operasi penghapusan (delete) dari kata yang terdeteksi typo dan mencarinya dalam kamus.[4] Proses operasi diterapkan untuk semua huruf pada kata yang salah ejaan secara bergantian. Setiap satu langkah dalam masing-masing proses dapat menghasilkan satu kata yang berbeda dari kata awal, kemudian kata-kata tersebut dicek dalam kamus daftar kata lalu disimpan, dan kemudian digunakan sebagai kandidat-kandidat kata.

2.4. Metode N-gram

Penggunaan N-gram telah banyak digunakan untuk berbagai masalah. Seperti prediksi kata, koreksi ejaan, pengenalan suara, koreksi kata terjemahan dan pencarian string. Metode N-gram mengambil potongan karakter kata dengan jumlah *n* dalam sebuah kalimat. N-gram dapat dibedakan berdasarkan sejumlah n, nilai n = 1 adalah unigrams, n = 2 adalah bigrams, n = 3 adalah Trigrams.[5]

$$P(W_n W_{n-1}) = \frac{P(W_{n-1} W_n)}{P(W_{n-1})} \quad (1)$$

Contoh pemenggalan kalimat metode N-gram dengan contoh kalimat “Covid masih berlangsung” sebagai berikut:

Unigrams : covid, masih, berlangsung
Bigrams : covid masih, masih berlangsung
Trigrams : covid masih berlangsung

2.5. Precision and Recall

Penggunaan *precision and recall* digunakan untuk menghitung pengujian pada aplikasi ini. Setiap hasil pengujian dikategorikan terlebih dahulu

Kategori dalam metode *precision and recall* adalah TP, TN, dan FN. *Precision* juga dapat digunakan untuk mengukur persentase metode dapat memberikan saran kata yang benar dari beberapa kandidat kata yang diberikan. *Recall* juga dapat digunakan untuk mengukur persentase metode akan memberikan saran kata yang benar dari beberapa jumlah kata yang sebenarnya. Akurasi digunakan untuk mengukur persentase total dari aplikasi ini.

$$Akurasi = (TP + TN)/(TP + TN + FP + FN) \quad (2)$$

$$Precision = \frac{TP}{TP + TN} \times 100\% \quad (3)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (4)$$

TP = Sistem menghasilkan data hasil koreksi benar
TN = Menghasilkan hasil koreksi data salah.
FN = Tidak menghasilkan hasil koreksi.
FP = Tidak adanya data dan tidak adanya hasil koreksi.

Apabila sistem menghasilkan koreksi yang benar, dikategorikan sebagai TP. Jika menghasilkan koreksi kata yang salah, akan dikategorikan sebagai TN. Jika tidak menghasilkan satu pun hasil koreksi kata, maka dikategorikan sebagai FN. Dan apabila tidak terdapat data dan tidak menghasilkan hasil koreksi, maka dikategorikan sebagai FP.

3. Hasil Pengujian

3.1 Pengujian Perhitungan Akurasi

Pengujian akurasi merupakan pengujian terhadap fitur tambahan pada aplikasi *Spelling Correction* ini yaitu pengujian dengan melakukan perhitungan akurasi dengan menggunakan hasil koreksi kata yang benar atau salah. Dari 55 dokumen yang ingin dicek dan dikoreksi masing-masing ada kata yang benar dikoreksi dan juga kata yang salah misalnya harusnya kata yang benar adalah "makan" setelah dikoreksi hasil koreksinya berubah menjadi "main" dan bukan menjadi kata yang benar setelah dikoreksi. Dari setiap dokumen yang sudah dicek dan dikoreksi mendapatkan beberapa hasil kata yang benar dan juga hasil kata yang salah, untuk mendapatkan hasil akurasinya dapat melakukan perhitungan *Precision and Recall.* Berikut ini adalah hasil pengujian akurasi dengan metode *precision and recall* dari kategori yang terdapat pada metode *precision and recall.* Dapat dilihat pada Tabel 1.

TABEL 1
Hasil Pengujian Akurasi

Kalimat	TP	TN	FN	Precision	Recall	Al
1	1					
2		1				
3	1					
4	1					
5		1				
6	1					
7	1					
8			1			
9	1					
10			1			
11	1					
12		1				
13	1					
14			1			
15			1			
16			1			
17	1					
18	1					
19	1					
20			1			
21	1					
22	1					
23			1			
24	1					
25		1				
26	1					
27		1				
28	1					
29			1			
30			1			
31		1				
32	1					
33	1					
34	1					
35	1					
36			1			
37	1					
38		1				
39			1			
40			1			
41	1					
42	1					
43			1			

Dokumen	TP	TN	FN	Precision	Recall	Akurasi
44	1					
45			1			
46	1					
47		1				
48		1				
49		1				
50			1			
51	1					
52		1				
53			1			
54	1					
55			1			
TOTAL	28	10	17	73.684%	37.037 %	69.09 %

Dari hasil pengujian diatas dapat dilihat total untuk TP sebanyak 28 kalimat, untuk TN sebanyak 10 kalimat dan untuk FN sebanyak 17 kalimat. Untuk nilai persentase *precision* adalah 73.684%, untuk nilai *recall* adalah 37.037%, dan untuk nilai akurasi adalah 69.09%. Dari hasil pengujian diatas kedua metode ini tidak dapat menghasilkan kata yang benar terhadap beberapa kasus kesalahan. Baik itu menghasilkan kata yang salah, maupun juga kata yang tidak menghasilkan satupun hasil koreksi kata. Hal ini dapat disebabkan oleh beberapa hal, yaitu:

1. Metode peter norvig tidak dapat mengoreksi kesalahan 2 huruf pada 1 kata. Sehingga sistem tidak dapat menganggap kata tersebut benar dan tidak dapat mengoreksinya.
2. Kata-kata di dalam korpus masih ada yang keliru dikarenakan banyak kata yang nilai probabilitas bigram nya sama maka sistem akan menghasilkan hasil koreksi katanya secara acak (random). Maka hasil koreksi kata tersebut memiliki 2 kemungkinan di antara benar dan salah.
3. Kandidat-kandidat kata yang dihasilkan metode Peter Norvig tidak terdapat didalam kamus kalimat dan tidak dapat dihitung nilai probabilitasnya dikarenakan tidak memiliki saran kata yang didapatkan dari kamus daftar kata.

3.2. Pengumpulan Data

Pengumpulan data dilakukan secara online dengan cara bertanya melalui media sosial kepada 10 mahasiswa dengan jumlah kalimat masing-masing 5 kalimat yang terdapat kata typo sehingga menjadi 50 dokumen dan pembuat aplikasi menambahkan 5 kalimat yang terdapat kata typo sehingga total terdapat 55 dokumen yang harus dikoreksi. Hasil koreksi dari pengumpulan data tersebut digunakan untuk perhitungan akurasi pada sistem ini. Berikut ini data pengujian dari 55 kalimat pada Tabel 2.

TABEL 2
Kalimat Pengujian

No	Kalimat	Kata *Typo*
1	ayah suka naik jerpah	jerpah
2	kerja keras bagai quda	quda
3	aku cinsa sama kamu	cinsa
4	plant bumi berwarna biru	plant
5	maaf kak, salah ketek	ketek
6	jaln kaki bersama aku	jaln
7	wku mau pergi ke mars	wku
8	slalu salah di mata dia	slalu
9	aku mau petgi renang	petgi
10	kamu jaht sekali ya	jaht
11	kita harus rajn olahraga	rajin
12	makan nais ayam	nasi
13	gigi saya berlbang	berlubang
14	rajin bermalas malasn	malasan
15	senang bersedih sdihan	sedihan
16	pintar mmbodoh bodohkn	membodoh
17	jangan maib game mulu	main
18	besik selasa guys	besok
19	Hari ini gas revsi	revisi
20	suapya nanti bisa nyantai	supaya
21	kaoan nih main pingping	kapan
22	aku dah punjam brt erwin nih	pinjam
23	bet erwin aqu pinjan unruk yanto	aku
24	ayok tin habus revisi maib	habis
25	ayah kh bukan ayah mu	ku
26	selamat datang slamat berbelanja	selamat
27	bolh kaka risolnya	boleh
28	jangan lupa tetp memakai masker	tetap
29	efektid membunuh kuman	efektif
30	untuk mewujudkan rulusan	lulusan
31	jangan lupa curi tangan	cuci
32	maap nomoer yang anda tuju sibuk	nomor
33	tidak mengampuni krang	orang
34	kenal kan nama sya	saya
35	lulus langsong kerja	langsung
36	bunga itu wangy sekali	wangi
37	jangn pernah menipu orang	jangan
38	ikut ibu ke pazar	pasar
39	jakarta ibukota negra	negara

	Indonesia	
40	memiliki hutanf harus dibayar	hutang
41	matri tersebut susah sekali	materi
No	Kalimat	Kata *Typo*
42	kapn kapan kita pergi ke Bandung	kapan
43	acara televisi seekarang membosankan	sekarang
44	serelah vaksin tetap menerapkan protokol kesehatan	setelah
45	apakah airr itu basah	air
46	jangan bercanda di saat gnting	genting
47	cuci tangan sebelum maka	makan
48	main sepat roda	sepatu
49	susa sekali naik motor	susah
50	kentang gorebg nikmat sekali	goreng
51	ibu gemae berpergian	gemar
52	hidup hanya skali	sekali
53	makan jangan berlebhan	berlebihan
54	hidup seht dan bermartabat	sehat
55	erwin selalu lapar setap saat	setiap

Dari 55 kalimat terdapat beberapa kalimat yang memiliki kesalahan 2 huruf pada 1 kata dimana metode peter norvig tidak dapat mengoreksi kesalahan ejaan seperti itu. Terdapat juga beberapa kalimat yang mengandung nama orang sehingga kata yang mengandung nama orang itu termasuk kesalahan ejaan karena kata tersebut tidak terdapat didalam kamus daftar kata KBBI.

4. Kesimpulan dan Saran

4.1 Kesimpulan

Berdasarkan hasil pengujian terhadap program *Spelling Correction* dengan menggunakan metode *peter norvig* dan *ngram* didapatkan kesimpulan sebagai berikut:

1. Aplikasi sudah dapat mengoreksi dokumen berupa text dan aplikasi sudah berhasil mengubah kata-kata yang mengandung salah
2. Pada hasil pengujian mendapatkan nilai akurasi hasil *Spelling Correction* sebesar 69.09% dengan menggunakan 55 dokumen sebagai data pengujiannya.
3. Jika aplikasi ini menggunakan lebih banyak data pengujian maka hasil akurasi pada aplikasi ini akan kurang dari 69.09% dikarenakan tidak adanya batasan kategori sehingga kata-kata yang terdapat dalam kamus kalimat tidak bisa mencangkup seluruh kata-kata umum yang terdapat dalam dokumen data penguji sehingga tidak dapat dikoreksi karena tidak adanya kata tersebut di dalam kamus kalimat.

4.2. Saran

Terdapat beberapa saran bagi yang ingin mengembangkan program Aplikasi Spelling Correction dengan menggunakan metode Peter Norvig dan N-Gram sebagai berikut:

1. Aplikasi sangat bergantung pada kualitas kamus daftar kalimat, sehingga kamus kalimat harus mencangkup banyak kalimat-kalimat umum.
2. Menambahkan beberapa fitur untuk mengoreksi Word dan PDF.
3. Mengoreksi kata yang memperhatikan frasa-frasa pada kalimat

REFERENSI

[1] Mutammimah; Sujaini, Herry; dan Nyoto, Rudy Dwi. "Analisis Perbandingan Metode Spelling Corrector Peter Norvig dan Spelling Checker BK-Trees pada Kata Berbahasa Indonesia (Studi Kasus: Universitas Tanjungpura)". Jurnal Sistem dan Teknologi Informasi. Vol. 5, Nomor 1. Pontianak: Fakultas Teknik Universitas Tanjungpura, 2017.

[2] Augusfian, Fendy; M. Viny, Christanti; Hendyli, Janson; dan Naga, Dali S. "Sistem Pengoreksian Ejaan Bahasa Indonesia dengan Damerau Levenshtein Distance dan Recurrent Neural Network (Studi Kasus: Universitas Tarumanagara)". Journal of Computer Science and Information Systems. 3/2, *hlm*. 144-152. Jakarta: Fakultas Teknologi Informasi Universitas Tarumanagara, 2019.

[3] Maghfira, Trusty Nadia; Cholissodin, Imam; dan Widodo, Agus Wahyu. "Deteksi Kesalahan Ejaan dan Penentuan Rekomendasi Koreksi Kata yang Tepat Pada Dokumen Jurnal JTIIK Menggunakan Dictionary Lookup dan Damerau-Levenshtein Distance (Studi Kasus: Universitas Brawijaya)". Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. Vol. 1, Nomor 6. Malang: Fakultas Ilmu Komputer Universitas Brawijaya, 6 Juni 2017.

[4] Gusdiwangsa, Rangga. "Perbaikan Kesalahan Ejaan Dengan Metode Symspell Pada Kasus

Tanya Jawab Dalam Bahasa Indonesia". Jurnal Universitas Komputer Indonesia. bab 2, h. 10. Bandung: Universitas Komputer Indonesia, 2019.

[5] Simajuntak, Maya Salinka; Sujaini, Herry; dan Safriandi, Novi. "Spelling Corrector Bahasa Indonesia dengan Kombinasi Metode Peter Norvig dan N-gram (Studi Kasus: Universitas Tanjungpura)". Jurnal Edukasi dan Penelitian Informatika. Vol. 4, Nomor 1. Kalimantan Barat: Universitas Tanjungpura, Juni 2018.

Ricky Martin, Seorang mahasiswa pada program studi Fakultas Teknologi Informasi di Universitas Tarumanagara.