Computatio: Journal of Computer Science and Information Systems, 3/2 (2019), 144-152

# SISTEM PENGOREKSIAN EJAAN TEKS BAHASA INDONESIA DENGAN DAMERAU LEVENSHTEIN DISTANCE DAN RECURRENT NEURA L NETWORK

**Fendy Augusfian [1], Viny Christanti M. [2], Janson Hendryli[3], Dali S. Naga [4]**
[1, 2, 3, 4] Program Studi Teknik Informatika, Fakultas Teknologi Informasi
Universitas Tarumanaraga
Jl. Letjen S Parman no 1, Jakarta 11440 Indonesia
*535150064.fendy@fti.untar.ac.id, viny@untar.ac.id, jansonh@fti.untar.ac.id, dalinaga@gmail.com*

***Abstrak***

*Tujuan dari penelitian ini adalah untuk membuat sistem pengoreksian ejaan teks Bahasa Indonesia, yang memiliki kemampuan untuk menangani dan memperbaiki kesalahan ejaan, baik kesalahan kata tidak sah maupun kesalahan kata sah. Sistem koreksi ejaan yang sudah ada dianalisis kembali dan dilakukan beberapa penyesuaian dan koreksi untuk meningkatkan akurasi. Sistem koreksi ejaan yang diusulkan dibuat dengan metode Damerau-Levenshtein, yang digunakan dengan penyesuaian dan koreksi dalam sistem koreksi ejaan yang sudah ada. Pencapaian yang dicapai oleh sistem koreksi ejaan yang sudah ada menghasilkan akurasi kata sebesar 40,6% dan kecepatan pemrosesan rata-rata 18,4 milidetik per kalimat dibandingkan hasil yang dicapai oleh sistem yang menggunakan Damerau-Levenshtein Distance dan Recurrent Neural Network Akurasi menghasilkan akurasi kata sebesar 21,3% dan kecepatan pemrosesan rata-rata adalah 29,21 milidetik per kalimat. Hasil pengujian ulang teks yang dicapai oleh sistem menggunakan Damerau-Levenshtein Distance dan Recurrent Neural Network menunjukkan akurasi kata sebesar dari 74%.*

***Kata kunci****—damerau-levenshtein distance, deep learning, teks bahasa Indonesia, n-gram, recurrent neural network*

***Abstract***

*This research was intended to create Indonesian Text Spelling Correction system with the capability to handle and make correction to both kind of spelling errors, non-word and real-word errors. Existing spelling correction system was analyzed and made some adjustment and modifications to boost its accuracy. The proposed spelling correction system is built with Damerau-Levenshtein Distance that used in existing spelling correction system along with the adjustment and modifications. The result that achieved by the system that uses by existing spelling correction with the word level accuracy of 40.6% and an average processing speed of 18.4 ms per sentence while the result that achieved by the system that uses Damerau-Levenshtein Distance and Recurrent Neural Network with the word level accuracy of 21.3% and an average processing speed of 29.21 ms per sentence. The result of retest text that achieved by the system that uses Damerau-Levenshtein Distance and Recurrent Neural Network with the word level accuracy of 74%*

***Keywords—****damerau-levenshtein distance, deep learning, Indonesian text, n-gram, recurrent neural network*

## 1. PENDAHULUAN

Bahasa adalah kemampuan yang dimiliki manusia untuk berkomunikasi dengan manusia lainnya menggunakan tanda, misalnya kata dan gerakan [1]. Ada 2 bentuk bahasa yaitu secara tertulis dan tidak tertulis [2]. Tujuan dari menulis itu sendiri adalah untuk menyampaikan makna yang terkandung di dalam kata. Kesalahan saat penulisan dapat membuat arti dari kata yang disampaikan menjadi keliru atau memiliki arti lain.

Pengoreksian ejaan adalah sistem yang dibuat untuk mendeteksi kesalahan ejaan dan memperbaikinya. Ada 2 macam 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 memiiki 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" [3].

Sistem pengoreksian ejaan yang sudah ada sebelumnya hanya dapat mendeteksi dan meperbaiki satu macam jenis kesalahan ejaan saja. Metode yang digunakan juga berbeda satu dengan yang lainnya. Setiap sistem pengoreksian ejaan menangani setiap kesalahan secara terpisah. Tujuan utama dari penelitian ini adalah untuk membuat sistem pengoreksian ejaan dengan kemampuan untuk memperbaiki kesalahan kata sah dan kesalahan kata tidak sah dengan tingkat akurasi kata yang lebih tinggi dibandingkan sistem pengoreksian ejaan sebelumnya. Penelitian relevan yang dijadikan sebagai referensi yaitu penelitian dari Rudy yang menggunakan Damerau Levenshtein Distance dan N-Gram [4]. Pada penelitian ini, Rudy menunjukkan penggunaan Damerau Levenshein Distance yang digabungkan dengan N-Gram mampu menangani sebagain dari kesalahan kata sah dan tidak sah dalam waktu yang cukup singkat.

Salah satu metode yang digunakan oleh Rudy yaitu Damerau Levenshtein Distance digunakan untuk pengembangan lebih lanjut dengan menggabungkannya dengan Recurrent Neural Network untuk meningkatkan akurasi kata dan performa dalam mendeteksi dan memperbaiki kesalahan kata. Recurrent Neural Network digunakan untuk menghasilkan perbaikan yang sesuai dengan data yang sudah latih disebelumnya. Sedangkan Damerau Levenshtein Distance digunakan untuk menghitung perubahan yang terjadi pada kata yang diperiksa dengan kata yang sesungguhnya dan meperbaikinya.

## 2. METODE PENELITIAN

Dalam penulisan kata, ada berbagai macam kesalahan baik kesalahan kata sah dan kesalahan kata tidak sah. Kesalahan tidak sah memiliki beberapa bentuk seperti deletion, insertion, substitution, transposition dan split word [5]. Pada penelitian ini, semua kesalahan kata sah dan kesalahan kata tidak akan diperbaiki menggunakan 2 metode yaitu Damerau Levenshtein Distance dan Recurrent Neural Network. Ilustrasi untuk kesalahan kata tidak sah dapat dilihat pada Tabel 1.

**Tabel 1** Ilustrasi kesalahan kata tidak sah

| Substitution | Deletion |
|---|---|
| Supstitution | Deltion |
| | |
| Insertion | Transposition |
| Insaertion | Transpositino |

2.1 Damerau Levenshtein Distance

Metode Levenshtein Distance adalah salah satu metode yang digunakan pada referensi sistem dan digunakan kembali pada penelitian ini. Levenhstein Distance adalah operasi untuk menghitung jumlah operasi yang diperlukan untuk mengubah suatu kata menjadi kata yang lainnya [6]. Levenshtein Distance menghitung operasi seperti deletion, insertion dan substitution sehingga cocok dalam pengoreksian ejaan yang dibuat [7]. Variasi dari Levenshein Distance adalah Damerau Levenshtein Distance, yang menambahkan variasi operasi transposition berdsarkan huruf yang berdekatan dan mengalami kesalahan ejaan [8]. Berdasarkan [6], [7] dan [8], rumus dari Levenshtein Distance dapat dilihat pada (1), (2), (3) dan Damerau Levenshtein Distance dapat dilihat pada (4).

$$d_{i0} = \sum_{k=1}^{i} w_{del}(b_k) \quad , \quad for\ 1 \le i \le m \tag{1}$$

$$d_{0j} = \sum_{k=1}^{j} w_{ins}(a_k) \quad , \quad for\ 1 \le j \le m \tag{2}$$

$$d_{ij} = \begin{cases} d_{i-1,j-1} & for\ a_j = b_i \\ min \begin{cases} d_{i-1,j} + w_{del}(b_i) \\ d_{i,j-1} + w_{ins}(a_j) \\ d_{i-1,j-1} + w_{sub}(a_j.b_i.) \end{cases} & for\ 1 \le i \le m, 1 \le j \le n \\ & for\ a_j \ne b_i \end{cases} \tag{3}$$

$$d_{ij} = \begin{cases} d_{i-1,j-1} & for\ a_j = b_i \\ min \begin{cases} d_{i-1,j} + w_{del}(b_i) \\ d_{i,j-1} + w_{ins}(a_j) \\ d_{i-1,j-1} + w_{sub}(a_j.b_i) \\ d_{i-2,j-2} + w_{tra} \end{cases} & if\ i,j>1, a_i = b_{j-1}, a_{i-1} = b_j \\ & for\ a_j \ne b_i \\ min \begin{cases} d_{i-1,j} + w_{del}(b_i) \\ d_{i,j-1} + w_{ins}(a_j) \\ d_{i-1,j-1} + w_{sub}(a_j.b_i) \end{cases} & for\ 1 \le i \le m, 1 \le j \le n \\ & for\ a_j \ne b_i \end{cases} \tag{4}$$

2.2. Recurrent Neural Network

Recurrent Neural Networks (RNN) adalah salah satu bagian dari jaringan saraf tiruan yang melakukan proses data secara sekuensial. RNN dapat memproses rangkaian yang berupa $x_1,\ldots,x_T$. RNN akan mengeluarkan rangkaian keluaran dengan jumlah elemen yang sama dengan jumlah elemen pada sekuens masukan [9].

RNN saling berbagi parameter pada bagian-bagian yang berbeda di dalam model. Berbagi parameter memungkinkan untuk memperpanjang dan menerapkan model pada bentuk-bentuk yang berbeda serta menyamakan mereka. Saling berbagi menjadi penting ketika ada suatu informasi yang dapat saja muncul pada berbagai macam posisi di dalam sekuens [9].

RNN yang digunakan pada penelitian ini adalah RNN Encoder-Decoder. RNN Encoder-Decoder adalah jaringan saraf buatan yang terdiri dari dua buah RNN. Encoder – Decoder biasanya digunakan untuk mengodekan rangkaian ke representasi vektor dengan panjang yang telah ditentukan sebelumnya. Vektor tersebut kemudian diterjemahkan kembali ke rangkaian yang lain sehingga panjang dari rangkaian dan masukkan dapat berbeda satu sama lain. Hal ini dapat digunakan untuk membuat sistem pengoreksian ejaan.

Encoder adalah jenis RNN yang membaca masukan rangkaian x secara berurutan. Hidden state dari RNN ini sendiri sama seperti menghitung RNN biasa. Hidden state terakhir c dari RNN merupakan representasi dari keseluruhan rangkaian masukan [10].

Decoder merupakan RNN yang dilatih untuk membuat rangkaian keluaran dengan memprediksi kata $y_t$ berikutnya berdsarkan hidden state $h_t$. Namun variabel $y_t$ dan $h_t$ dipengaruhi

oleh $y_{t-1}$ dan hasil c dari rangkaian masukan [10]. Oleh karena itu, hidden state dari decoder pada waktu t adalah: $ht = f(h_{t-1}, y_{t-1}, c)$ (5) dengan f sebagai fungsi aktivasi.

Algoritma untuk perhitungan RNN Encoder-Decoder adalah sebagai berikut:

1. Menentukan *Hidden State* (c).
2. Lakukan propagasi maju *Encoder* hingga t = jumlah huruf sekuens – 1 dengan:
   a. Inisialisasi matriks U berukuran nilai v dimana berisi jumlah kolom pada matriks X dikali nilai dari *hidden state*(c) yang berisi nilai acak dengan interval $\left[\frac{-1}{\sqrt{v}}, \frac{1}{\sqrt{v}}\right]$.
   b. Inisialisasi matriks W berukuran nilai dari *hidden state* (c) dikali nilai *hidden state* (c) yang berisi nilai acak dengan interval $\left[\frac{-1}{\sqrt{c}}, \frac{1}{\sqrt{c}}\right]$.
3. Lakukan propagasi maju *Decoder* hingga t = t sebelumnya + jumlah huruf keluaran – 1 dengan:
   a. Inisialisasi matriks U berukuran nilai v dimana berisi jumlah kolom pada matriks X dikali nilai dari *hidden state*(c) yang berisi nilai acak dengan interval $\left[\frac{-1}{\sqrt{v}}, \frac{1}{\sqrt{v}}\right]$.
   b. Inisialisasi matriks V berukuran nilai dari *hidden state* (c) dikali jumlah huruf pada hasil keluaran yang berisi nilai acak dengan interval $\left[\frac{-1}{\sqrt{c}}, \frac{1}{\sqrt{c}}\right]$.
   c. Inisialisasi matriks W berukuran nilai dari *hidden state* (c) dikali nilai *hidden state* (c) yang berisi nilai acak dengan interval $\left[\frac{-1}{\sqrt{c}}, \frac{1}{\sqrt{c}}\right]$.
4. Lakukan propagasi mundur Decoder hingga t = t sebelumnya - jumlah huruf keluaran.
5. Lakukan propagasi mundur Encoder hingga t = 0
6. Matriks U, V dan W pada Decoder dan matriks U dan W pada Encoder akan dijadikan model untuk melakukan pengoreksian ejaan.

### 2.3 Data yang digunakan

Data yang digunakan dalam penelitian ini adalah kamus Bahasa Indonesia, koleksi dokumen artikel berita, dan teks pengujian. Kamus Bahasa Indonesia digunakan sebagai pembanding dalam metode Damerau Levenshtein Distance. Koleksi dokumen artikel berita digunakan sebagai bahan training dari metode RNN yang digunakan. Sedangkan teks pengujian dipakai untuk menguji hasil dari metode metode yang telah ditentukan.

#### 2.3.1 Kamus Bahasa Indonesia

Data yang diperlukan untuk penelitian ini salah satunya adalah kamus Bahasa Indonesia. Kamus Bahasa Indonesia didapatkan dari website indodic.com. Pada website ini, kamus dapat langsung diunduh pada link yang telah disediakan. Kata yang terdapat dalam kamus ini berjumlah 41312 kata. Namun, kamus ini belum mencakup keseluruhan dari kata Bahasa Indonesia yang ada sehingga harus dilakukan penyuntingan kembali.

#### 2.3.2 Koleksi Dokumen Artikel Berita

Koleksi dokumen untuk penelitian ini didapatkan dari alumni mahasiswa UNTAR yaitu Rudy. Koleksi dokumen yang digunakan adalah dokumen artikel media berita online yaitu Kompas. Bagian dokumen yang digunakan untuk training pada RNN adalah judul dan isi berita. Dokumen artikel berita berupa file XML dengan character encoding UTF-8.

Jumlah dari dokumen artikel media berita online adalah 5000 berita yang dibagi kedalam 5 file berbeda. Dokumen kemudian dipindahkan kedalam format RTF untuk dapat digunakan sebagai bahan training. Rincian kategori dokumen berita yang digunakan dapat dilihat pada Tabel 2.

**Tabel 2** Statistik Kategori Berita pada Dokumen Training

| Kategori | Jumlah |
|---|---|
| Regional | 867 |
| Megapolitan | 83- |
| Nasional | 714 |
| Entertainment | 393 |
| Bola | 365 |
| Ekonomi | 355 |
| Internasional | 294 |
| Otomotif | 228 |
| Travel | 204 |
| Properti | 184 |
| Olahraga | 142 |
| Tekno | 140 |
| Health | 136 |
| Female | 99 |
| Sains | 34 |
| Edukasi | 15 |

2.3.3 Kamus Bahasa Indonesia

Teks pengujian didapatkan dari alumni Mahasiswa UNTAR yaitu Rudy yang berisi 50 kalimat dalam 3 bentuk berbeda. Bentuk yang pertama adalah kalimat yang memiliki kesalahan tidak sah. Bentuk yang kedua adalah kalimat yang memiliki kesalahan sah dan yang terakhir adalah gabungan dari keduanya. Total dari kalimat yang diuji adalah 150 kalimat yang terdiri dari 50 pada setiap bentuk kalimat.

## 3. HASIL DAN PEMBAHASAN

3.1 Referensi Sistem

Hasil percobaan dari referensi system dapat dilihat pada Tabel 3.

**Tabel 3** Hasil percobaan Referensi Sistem

| Kategori Pengujian | Sah | Tidak Sah | Gabungan |
|---|---|---|---|
| **Akurasi Kalimat** | 36 % | 44 % | 24 % |
| **Akurasi Kata** | 37 % | 46 % | 39 % |
| **False Postive** | 0 % | 0 % | 0 % |
| **Kecepatan Rata Rata (ms)** | 20.7 | 18.8 | 15.7 |
| **Akurasi Kata Rata-Rata** | (46 + 37 + 39 / 3) = 40.6 % | | |

Berdasarkan dari hasil percobaan, didapatkan bahwa referensi sistem dapat menghasilkan akurasi yang cukup seimbang antara teks pengujian kata tidak sah dan sah serta kecepatan pemrosesan yang terbilang cepat. Referensi sistem juga dapat memperbaiki kalimat yang termasuk dalam kesalahan kata sah dan tidak sah dengan akurasi yang cukup baik dengan metode Levenshtein Distance dan N-Gram yang diterapkan dibandingkan dengan metode Recurrent Neural Network yang digunakan pada sistem yang akan dibuat.

Namun dalam mendeteksi kesalahan kata tidak sah, sistem pengoreksian ejaan ini memiliki akurasi yang rendah dibandingkan dengan metode sebelumnya yang digunakan pada sistem pengoreksian ejaan yang dibuat yaitu metode Damerau Levenshtein Distance.

Hasil percobaan dari meteode Damerau Levenshtein Distance secara manual dan otomatis dapat dilihat pada Tabel 4 dan 5 sedangkan hasil percobaan dari metode Recurrent Neural Network dapat dilihat pada Tabel 6.

**Tabel 4** Hasil percobaan dengan Damerau Levenshtein Distance Manual

| **Kategori Pengujian** | Sah | Tidak Sah | Gabungan |
|---|---|---|---|
| **Akurasi Kalimat** | 0 % | 88 % | 0 % |
| **Akurasi Kata** | 0 % | 84 % | 42 % |
| **False Postive** | 0 % | 0 % | 0 % |
| **Kecepatan Rata Rata (s)** | 4.6 | 4.3 | 5.2 |

**Tabel 5** Hasil percobaan dengan Damerau Levenshtein Distance Otomatis

| **Kategori Pengujian** | Sah | Tidak Sah | Gabungan |
|---|---|---|---|
| **Akurasi Kalimat** | 0 % | 70 % | 0 % |
| **Akurasi Kata** | 0 % | 71 % | 36 % |
| **False Postive** | 0 % | 0 % | 0 % |
| **Kecepatan Rata Rata (s)** | 2.1 | 2.1 | 2.8 |

**Tabel 6** Hasil percobaan dengan Recurrent Neural Network

| **Kategori Pengujian** | Sah | Tidak Sah | Gabungan |
|---|---|---|---|
| **Akurasi Kalimat** | 2 % | 8 % | 8 % |
| **Akurasi Kata** | 1 % | 25 % | 14 % |
| **False Postive** | 15 % | 14 % | 15 % |
| **Kecepatan Rata Rata (s)** | 18.5 | 18.1 | 19.2 |

3.2 Sistem yang dibuat

Sistem dibuat dengan menggabungkan metode Damerau Levenshtein Distance dan Recurrent Network agar dapat memperbaiki kesalahaan kata sah maupun tidak sah dengan akurasi yang cukup baik. Hasil dari percobaan sistem dapat dilihat pada Tabel 7.

**Tabel 7** Hasil percobaan dengan sistem yang dibuat

| **Kategori Pengujian** | Sah | Tidak Sah | Gabungan |
|---|---|---|---|
| **Akurasi Kalimat** | 10 % | 20 % | 6 % |
| **Akurasi Kata** | 9 % | 36 % | 19 % |
| **False Postive** | 13 % | 12 % | 14 % |
| **Kecepatan Rata Rata (s)** | 29.2 | 28.2 | 29.4 |
| **Akurasi Kata Rata-Rata** | (36 + 9 + 19 / 3) = 21.3 % | | |

Dari hasil percobaan, dapat disimpulkan bahwa metode gabungan memiliki akurasi yang berada diantara kedua metode yang dipakai. Hal ini dikarenakan metode gabungan tetap dapat memperbaiki kesalahan kata sah maupun tidak sah dibandingkan metode RNN saja karena dibantu oleh metode Damerau Levenshtein Distance.

Metode Damerau Levenshtein Distance membantu untuk memperbaiki kesalahan kata tidak sah yang tidak dapat diperbaiki oleh metode RNN. Namun false-positive yang terjadi berada pada tingkatan yang hampir sama dengan metode RNN dikarenakan ketidakmampuan DLD untuk memperbaiki false-positive. Oleh karena itu, dilakukan pengujian ulang dengan 20 kalimat yang memiliki berbagai bentuk kesalahan yang dibuat untuk mengetahui kemampuan metode Gabungan untuk memperbaiki kesalahan kata tidak sah, sah maupun kesalahan gabungan Hasil dari percobaan ulang metode Gabungan dapat dilihat pada Tabel 8.

Keterangan:

E1 = Transposition Error
E2 = Insertion Error
E3 = Deletion Error
E4 = Substitution Error
E5 = Split Word Error
E6 = Real Word Error
E7 = Gabungan Error

**Tabel 8** Hasil percobaan ulang metode Gabungan.

| Kategori Pengujian | E1 | E2 | E3 | E4 | E5 | E6 | E7 |
|---|---|---|---|---|---|---|---|
| **Akurasi Kalimat** | 0 | 0 | 0 | 0 | 0 | 33 % | 0 % |
| **Akurasi Kata** | 85 % | 75 % | 75 % | 75 % | 66 % | 25 % | 57 % |
| **False Positive** | 0 % | 15 % | 16 % | 16 % | 20 % | 10 % | 8 % |
| **Kecepatan Rata-Rata (s)** | 16,3 | 15,9 | 16,1 | 15,4 | 15,7 | 15.8 | 15,7 |
| **Akurasi Kata Rata-Rata** | (85+75+75+75+66+25+57)/ 7 = 74 % | | | | | | |

Dari hasil pengujian ulang, dapat disimpulkan bahwa metode Gabungan dapat memperbaiki kesalahan ejaan tidak sah dalam berbagai bentuk dengan akurasi rata rata 74% namun belum dapat memperbaiki kesalahaan ejaan sah. Kesalahan ejaan tidak sah yang diperbaiki paling banyak yaitu transposition error. Sedangkan kesalahan ejaan tidak sah yang kurang dapat diperbaiki Split Word Error. False positive yang didapat yaitu dibawah 21 % meskipun kalimat belum dapat dikoreksi sepenuhnya. Hal ini membuktikan bahwa rendahnya akurasi pada pengujian sebelumnya dapat disebabkan oleh kejarangan data yang sesuai dengan kalimat yang dikoreksi.

3.3 Pembahasan

Berdasarkan hasil pengujian di atas, hasil dari metode Gabungan yang dibuat belum dapat menghasilkan hasil yang baik daripada metode yang digunakan oleh referensi sistem. Dari segi akurasi kata dan kalimat, metode gabungan yang dibuat tidak dapat melebihi akurasi metode yang digunakan pada referensi sistem di semua macam jenis kesalahan kata. Kecepatan dan lama pemrosesan dalam pengoreksian pun terbilang cukup lambat dibandingkan pada referensi sistem. Selain itu, juga dapat dibuktikan dengan hasil koreksi metode Gabungan yang lebih rendah dari referensi sistem yang diuji dengan hipotesis statistika.

Hal ini dapat terjadi karena metode Gabungan yang dibuat memiliki kendala dalam pengoreksian menggunakan salah satu metode yaitu metode RNN. Koreksi yang dilakukan oleh metode RNN masih banyak kekurangan seperti nilai false positive yang cukup besar, lama pemrosesan yang cukup lambat dan kesalahan kata sah yang belum dapat diperbaiki secara tepat. Selain itu, jika kalimat yang dikoreksi melebihi atau kurang dari jumlah kata tertentu memiliki kemungkinan besar metode RNN menghasilkan false positive yang cukup besar dan tidak dapat mengoreksi ejaan secara tepat. Hal ini membuat metode Damerau Levenshtein Distance sebagai salah satu metode gabungan tidak dapat memberikan dampak yang cukup besar pada pengoreksian karena metode Damerau Levenshtein Distance hanya dapat memperbaiki kesalahan kata tidak sah saja.

Namun pada pengujian ulang metode Gabungan yang sudah dilakukan sebelumnya, metode RNN yang digabungkan dengan Damerau Levenshtein Distance dapat memperbaiki kalimat dengan akurasi yang cukup baik. Dari hasil pengujian ulang, dapat diketahui bahwa metode Gabungan belum dapat memperbaiki secara tepat kesalahan kata tidak sah pada bentuk Insertion Error sedangkan untuk kesalahan lainnya dapat diperbaiki dengan cukup baik. Kalimat yang dapat dikoreksi oleh metode Gabungan harus terkandung pada kalimat yang berada pada

proses training. Kurangnya intensitas kata yang muncul pada kalimat dalam proses training dapat mengakibatkan ketidakmampuan metode Gabungan untuk memperbaiki ejaan.

Berdasarkan hal tersebut, dapat disimpulkan metode RNN hanya dapat mengoreksi kalimat yang sudah pernah dilatih sebelumnya pada saat training dilakukan. Sedangkan metode Damerau Levenshtein Distance dapat berfungsi dengan baik jika diterapkan pada kalimat yang dapat dikoreksi dengan cukup baik oleh metode RNN.

## 4. KESIMPULAN

Hasil dari penelitian dalam pembuatan sistem pengoreksian ejaan teks Bahasa Indonesia yang dapat memperbaiki kesalahan kata sah maupun tidak sah belum dapat menghasilkan akurasi yang baik dibandingkan dengan referensi sistem. Hasil yang didapat oleh sistem yang dibuat dengan menggunakan Damerau Levenshtein Distance dan Recurrent Neural Network yaitu dengan rata rata akurasi kata sebesar 21.3 % dan kecepatan rata rata sebesar 29.2 s.

Hal-hal yang dapat disimpulkan dari penelitian yang sudah dilakukan adalah:

1. Koreksi kesalahan eja kata bahasa Indonesia melalui metode koreksi gabungan Damerau Levenshtein Distance dan Recurrent Neural Network berhasil mengoreksi 44 kata dengan perbandingan referensi sistem yang berhasil mengoreksi 84 kata.
2. Koreksi kesalahan eja kata bahasa Indonesia melalui metode koreksi gabungan Damerau Levenshtein Distance dan Recurrent Neural Network memiliki akurasi kata rata-rata sebesar 21 % dan kecepatan rata rata 28.9 s dengan perbandingan referensi sistem yang memiliki akurasi kata rata-rata sebesar 41 % dan kecepatan rata-rata 18.4 ms.
3. Pada pengujian ulang, koreksi kesalahan eja kata bahasa Indonesia melalui metode koreksi gabungan Damerau Levenshtein Distance dan Recurrent Neural Network memiliki akurasi kata rata-rata sebesar 74 %.
4. Metode Recurrent Neural Network dapat menutupi kekurangan metode Damerau Levenshtein Distance karena dapat memperbaiki kesalahan ejaan sah yang tidak dapat diperbaiki sebelumnya namun banyak terjadi false positive saat melakukan pengoreksian.
5. Penggunaan metode Damerau Levenshtein Distance pada sistem pengoreksian ejaan gabungan dapat meningkatkan akurasi kalimat dan kata pada teks pengujian yang memiliki kesalahan ejaan tidak sah.

Saran yang dapat dipergunakan untuk pengembangan selanjutnya adalah dengan

1. Mengurangi waktu untuk melakukan pemrosesan baik menggunakan metode Damerau Levenshtein Distance, RNN maupun metode gabungan. Hal ini dapat dilakukan dengan menerapkan Finite State Automata dan struktur data trie agar dapat melakukan proses lebih cepat seperti yang sudah diterapkan oleh referensi sistem.
2. Menambahkan jumlah dokumen training yang telah divalidasi dengan meminta bantuan pakar. Hal ini dapat dilakukan dengan bekerja sama dengan sebuah lembaga bahasa namun akan memakan biaya dan waktu yang cukup lama.
3. Kamus Bahasa Indonesia yang digunakan juga harus divalidasi dan sesuai dengan kamus Bahasa Indonesia yang dipakai dan diterapkan di Indonesia sehingga kata yang dijadikan saran kata dapat sesuai.
4. Menggunakan dokumen training yang bukan berupa berita melainkan berasal dari buku berbahasa Indonesia, novel Bahasa Indonesia maupun subtitle Bahasa Indonesia dari film.

## DAFTAR PUSTAKA

[1] ZonaReferensi, "Pengertian Bahasa Menurut Para Ahli dan Secara Umum", https://www.zonareferensi.com/ pengertian-bahasa/, 18 Agustus 2018
[2] Pakar Komunikasi, "30 Cara Berkomunikasi dengan baik yang efektif", https://pakarkomunikasi.com/cara-berkomunikasi-dengan-baik, 19 Agustus 2018.
[3] Pedler, Jennifer. "Computer Correction of Real-word Spelling Errors in Dyslexic Text", (Birkbeck: London University)
[4] Rudy, "Sistem Pengoreksian Ejaan Untuk Bahasa Indonesia dengan Metode N-Gram dan Edit Distance", (Jakarta: Fakultas Teknologi Informasi Universitas Tarumanagara, (Skripsi tidak dipublikasikan), 2012.
[5] Fahma, Arina Indana.: Cholissodin, Imam.: Perdana, Rizal Setya. "Identifikasi Kesalahan Penulisan Kata (Typographical Error) pada Dokumen Berbahasa Indonesia Menggunakan Metode N-gram dan Levenshtein Distance". Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer. Vol.2, No 1. Malang: Fakultas Ilmu Komputer, Januari 2018.
[6] Levenshtein, "How Levenshtein Works", http://www .levenshtein.net/index.html, 28 Agustus 2018
[7] Levenshtein, Vladimir. "Binary codes capable of correcting deletions, insertions, and reversals", Moskow: Soviet Physics Doklady, 1966.
[8] Damerau, F. "A Technigque for Computer Detection and Correction of Spelling Errors". (New York: Cornell University, 1964).
[9] Goodfeloow, Ian. "Deep Learning", (Massachusetts: MIT Press, 2016)
[10] Cho, Kyunghyun. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing(EMNLP), (Oktober 2014)