CORRECTION SPELLING ERROR ON CONVERSION RESULTS FILE IN INFORMATION EXTRACTION

Abdul Rohim[1], Ken Kinanti Purnamasari[2]

[1,2] Informatic Engineering-Universitas Komputer Indonesia
Jl. Dipatiukur 112-114 Bandung
Email : rohimabdul531@gmail.com[1], ken.kinanti@email.unikom.ac.id[2]

ABSTRACT

Files generated from the file conversion process can cause spelling mistakes including non-word and word-boundary misspellings. Non-word spelling mistakes because spellings are not available in dictionaries so spellings do not match Indonesian standard of grammar (EYD). Non-word errors are dealt with using two spelling correction methods that are soundex methods to provide remedial remedies and methods of damerau-levenshtein distance used to select improvements from words recommended. The soundex method is a spelling correction technique (phonetic string matching) and damerau-levenshtein distance method based on the approximate string matching. Soundex will look for words that have similarities based on the same phonetic code as the recommended word and dread-levenshtein distance look for the smallest edit distance value of the word that is considered misspelled with the recommended word. While the word-boundary spelling mistake is because the spacing does not fit between the words so the two words become one word. Word-boundary errors are handled by looking for two words in KBBI, when the two words are combined will be the same as the word that is considered a word-bound misspelling. Based on test results, misspelled word detection have average of recall by 59.5% and precision by 70.6%, while word improvements have average of recall by 100% and precision by 71.4%.

Keyword : correction spelling error, soundex, damerau-levensthein distance, information extraction, non-word, word-boundary.

1. INTRODUCTION

File conversion is a way of converting from one file format to another. Converted files are usually a .pdf formatted document that is converted to either .doc or .docx format. This is usually done so that the information contained in the .pdf document can be retrieved because the .pdf document can be a collection of hard file scans. Therefore during the file conversion process a conversion error may occur that causes a misspelling of the word.

Word spelling mistakes are divided into two types: non-word spelling errors that occur when the wrong word is not in the dictionary [1][2][3] and real-word spelling errors that occur when the wrong word is in the dictionary, but not the word in question [1]. But in the spelling error correction research on file conversion results there is a word-boundary spelling mistake that occurs due to uneven spacing for each word[4]. Word-boundary spelling errors can occur on any word contained in a dictionary or a word not contained in the dictionary.

2. CONTENT OF STUDY

2.1 Related Work

Several studies have been done to detect and correct spelling mistakes using either approximate string matching techniques or based on phonetic string matching. As did Utis Sutisna and Julio Adisantoso[5] with approximate string matching technique using the damerau-levenshtein distance method to correct spelling of Indonesian query. The study resulted in an increase in precision in the retrieval performance by 22% from a previous study conducted by Arumsari [6] on spelling correction in Indonesian using levensthein method. The phonetic string matching technique as performed by Fahmilu Kurniawan [7] compares the soundex, phonix, metaphone and caverphone methods for word correction recommendation in the case of Indonesian spelling correction that yields the most accuracy of 92% for the soundex method.

2.2 Spelling Error

Spelling error is an error that occurs in a word so that the word is not in accordance with the Enhanced Spelling (EYD) or the change of meaning in a sentence. Spelling errors can be divided into 3 parts, that is:

1. Non-word spelling error
   The non-word spelling error is a word error where the word is not in the dictionary [1][2][3]. An example of a scholar who became a saijana, sajana.
2. Real-word spelling error
   The real-word spelling error is a mis-word where the wrong word is in the dictionary, but not the word in question [8][9][1]. Real-word errors usually disrupt the syntax and semantics of a sentence, so it takes human help to correct it. Example: I eat kunigan rice.

3. Word-boundary spelling error

   A word-bound error is a mis-word where the wrong word is caused by an inappropriate spacing placement [4]. Examples of eating habit become eatinghabit

2.3 Information Extraction

Information extraction is the retrieval of facts and structured information from large text collections. The notions of fact here are the various entities that are taken into account. Briefly the extraction of information is a process of obtaining the structured facts from the available data [10].

2.4 Rule-Based System

Rule based systems in this research are used at the time of extraction of information for ploting data to be processed and character modification based on change list and character change based on grammar rules. The input data plotted include the cover, the abstract. The change of characters based on grammatical rules is done on consonant letters based on consonant phonemes as follows.

2.5 Soundex

Soundex method is one of phonetic algorithms (based on speech similarity) and is the oldest algorithm. Soundex patented by Robert C. Russell in 1918. Soundex encodes the word into a soundex phonetic code into four digits. The first digit is the earliest letter in a word and then followed by three digit numbers obtained from the soundex rule. Soundex splits letters into seven different classes [12] as shown in Table III.2.

Tabel II. 1 Kode numerik metode *soundex* [12]

Code	Alphabet
No Coded	A, I, U, E, O, H, W, Y
1	B, F, P, V
2	C, G, J, K, Q, S, X, Z
3	D, T
4	L
5	M, N
6	R

2.6 Damerau-Levensthein Distance

This method was devised by Fred J. Damerau in 1964 describing the comparison of words with regard to four kinds of error [13] (given the example of the word ALPHABET) as exemplified by damerau in Ulrich Pfeifer, Thomas Poersch and Norbert Fuhr [14], that is:

1. Insertion of letters (insertion), into ALLPHABET.
2. Deletion of letters in words (deletion), into ALHABET.
3. Replacement of a letter with another letter (substitution), into ALPHIBET.
4. Redemption of two consecutive letters (transposition), into ALHPABET.

Here is a formula to find the value of edit distance on each recommended word.

$$d[i,j] = min\begin{cases} d[i-1,j]+1 \\ d[i,j-1]+1 \\ d[i-1,j-1]+cost \end{cases} \quad \dots 2.1$$

$$d[i,j] = min\{\, d[i,j], d[i-2,j-2] + cost \quad \dots 2.2 \,\}$$

Kemudian untuk mencari nilai jarak maksimum dari kata yang dianggap salah eja sebagai berikut.

$$jarak_max = round((\frac{40}{100}) \times M)$$

2.7 Dictionary

In this study Indonesian dictionary used is the result of research Rizki Hilmi Fauzi [15] because it contains enough words. The total words contained in the dictionary reached 52.978 words while the English dictionary uses a dictionary from Cambridge [16].

This dictionary will be used for the process of checking the word, if the input word is in the dictionary then the word is considered true and if not then the word is included into the spelling error of non-word or word-boundary.

2.8 System Stages

Previous studies used spell-proofing methods based on similarities in pronunciation or similarity in writing. In my research the method used in the process of detection and correction of spelling errors is the soundex and damerau-levensthein distance method. The whole process is divided into four main sections, namely information extraction, preprocessing, detection and correction of spelling errors, and mapping.

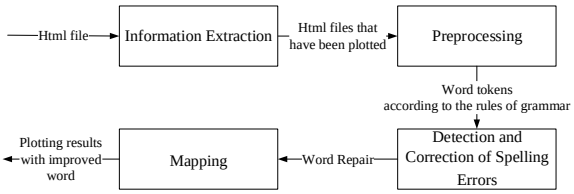


Figure II. 1 System Stages

a. Information Extraction

Stages of information extraction are divided into two stages, that is :

- Deletion of html entity characters, ie will be the process of removal of html entity characters which is the stage to remove the character of the space entity ie & nbsp; contained in the html file. This happens because spaces in html will only be executed once, even if you write a space ten times.
- Plotting is used to obtain structured data from unstructured input documents so that the data obtained matches the need.

b. Preprocessing

The preprocessing stage is the stage to prepare the input data from the extraction of information to be processed in the process of checking and correction of spelling errors. Preprocessing in this study consists of several stages, that is:

- Removal of html tags, ie removal of all html tags on plotting results such as tags * *, **, **, * *, **, and tag **.
- Character changes based on change list, Changing the character in question is a change that only occurs in a letter so that a word does not match the rules of spelling are justified. Changes of character letters that occur is a change from one character to another character such as a number, punctuation, and symbols even become characters of other letters.

Tabel II. 2 Character Changing Rules

Character On Document	Correction Result
IVI	M
], [	l
1>	D
<l	d

- Scanners are the process for grouping text into tokens whether they are letter characters, numbers, or symbols. This process is done to save the original text character form that will be used again during the mapping process to repair the text with the words that have been repaired based on the error that occurred at once will be used for the next process such as filtering, case folding, named entity tagger, and character change based on grammar rules as well as the process of detection and correction of spelling errors.
- Filtering, ie the deletion stage of all punctuation except point punctuation (.) Because it will be used in the Named Entity Tagger stage to detect the tag person so that at the time of checking the error the word human name will not be considered misspelled.
- Case folding, ie the stage to uniform all the characters into lowercase letters. This is done because at the stage of changing the characters based on the rules of grammar using lowercase as a rule.
- NE tagger (Named Entity), which is the stage to label a string whether the string is a human name or not. If the string is a human name it will be labeled person so that the string will not be compared with the dictionary to know the error he said. Named Entity detection using NER (Named Entity Recognition) library results from Yusuf syaifudin.
- Changing characters based on grammar rules, ie checking all tokens whether there is a consonant phoneme error or not. If there is a phoneme error it will be corrected in accordance with the rules of phoneme kosonan that exist in the Indonesian language. Here are the rules of phoneme error cosonan in this research.

Tabel II. 3 Kesalahan Fonem

Phoneme Error	Example	Improved results
/rn/	rnakan	makan
	rnasalah	masalah
	kearnanan	keamanan
	dalarn	dalam
/ch/	olch	oleh
/dm/	dmu	ilmu
/nn/	detenninasi	determinasi
/cn/	elemcn	elemen
/dc/	ordc	orde
	dcsa	desa
/lc/	lcbih	lebih
/sc/	sclesai	selesai
/cj/	cjata	data

c. Detection And Correction Spelling Errors

This stage consists of several stages for the process of detection and correction of spelling errors, that is:

- Determination of errors

Determination of word error is by comparing each word with word list contained in KBBI, dictionary acronym and cambridge dictionary. Words that are not detected in KBBI, an acronym dictionary and a cambridge dictionary are considered misspelled words. The process of determining the word error can be seen in figure III.2.

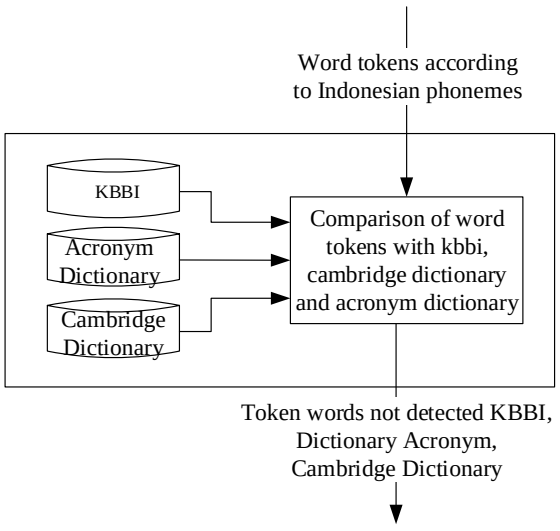


Figure II. 2 Word Error Determination Process

- Determination and Correction of Word-Boundary Spelling Errors

After the process of determining the error word then the wrong word list will then be checked whether from the word list there is a word-boundary spelling error. Word-boundary determination is checking word with word list contained in KBBI dictionary based on sequence of letters in every word token, if one of the word or all words checked detected KBBI then the word includes word-boundary error.

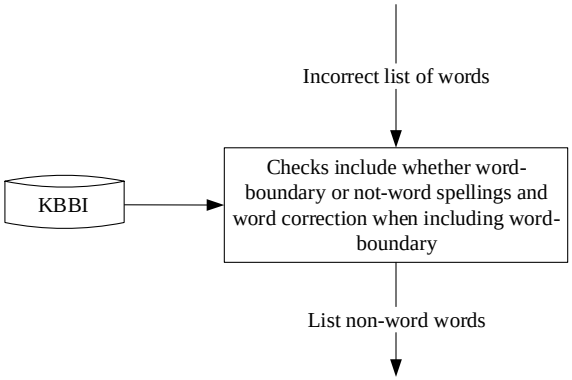


Figure II. 3 Word-Boundary Error and Correction Determination

The following is the path of detection and correction of word-boundary spelling errors.

- Defining Non-Word Spelling Recommendations

The process of determining word recommendations using the soundex method. Untested tokens when word-boundary checking will be considered non-word spelling errors. These tokens will be converted into phonetic code with phonetic code soundex rules. The recommendation is based on the similarity of the phonetic code between words considered to be misspelled non-word with a list of soundex phonetic codes in the dictionary. The following is the word recommendation process as shown in Figure III.5 below.

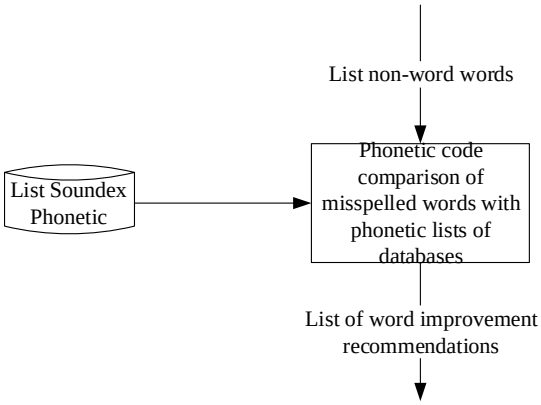


Figure III. 4 Recommendation Process for Non-Word Fixed Word Errors

For example a misspelled word is the word "disertasl". Then converted into soundex soundonic phonetic code into d263. The following is a recommendation for the word "disertasl".

Table II. 4 Word Recommendation

The misspelled word	The Phonetic Code	Recommendation
disertasl	d263	degradasi
		dekoratif
		dekorator
		dekret
		desertir
		diakritik
		disertasi
		diskredit
		disertai

- Determination of Selection of Non-Word Spelling Error Word

This stage is a selection process for word correction of recommended words. The selection of words for correction using the damerau-levenshtein distance method is to find the least edit distance value of the recommended words and the edit distance value is less than equal to the maximum distance value of the word considered misspelled. The following is the damerau-levenshtein algorithm.

1. Calculate the maximum distance from words that are considered misspelled non-word. It is assumed that the misspelled word is "disertasl". Then, M = 9

$$jarak_max = round((\frac{40}{100}) \times M)$$

$$jarak_max = round((\frac{40}{100}) \times 9)$$

$$jarak_max = 4$$

2. Then initialize s = "dissertation" along 9 characters and t = "disertasl" along 9 characters. Create an d matrix with rows of 0..9 and columns of 0..9.
3. Next compare each character to the matrix d. Starting from cell d [1,1], compare the character "d" on the line with the character "d" in the column. First determine the cost value. Because "d" = "d", then cost = 0. Then calculate the minimum value. Then get it

$$d[i,j] = min\begin{cases} d[i-1,j]+1 \\ d[i,j-1]+1 \\ d[i-1,j-1]+cost \end{cases}$$

$$d[1,1] = min\begin{cases} d[1-1,1]+1 = d[0,1]+1 = 1+1 = 2 \\ d[1,1-1]+1 = d[1,0]+1 = 1+1 = 2 \\ d[1-1,1-1]+cost = d[0,0]+0 = 0+0 = \mathbf{0} \end{cases}$$

$$d[1,1] = \mathbf{0}$$

Then, the matrix for d [1,1] is 0.

4. Selanjutnya untuk sel d[1,2], bandingkan Next to cell d [1,2], compare the character "i" with "i". because "i" = "i", then cost = 0. Then calculate the minimum value, obtained

$$d[i,j] = min\begin{cases} d[i-1,j]+1 \\ d[i,j-1]+1 \\ d[i-1,j-1]+cost \end{cases}$$

$$d[1,2] = min\begin{cases} d[1-1,2]+1 = d[0,2]+1 = 2+1 = 3 \\ d[1,2-1]+1 = d[1,1]+1 = 0+1 = \mathbf{1} \\ d[1-1,2-1]+cost = d[0,1]+0 = 1+0 = \mathbf{1} \end{cases}$$

$$d[1,2] = \mathbf{1}$$

Then, the matrix for d [1,2] is 1.

5. Perform for each matrix cell as in the second and third steps so that the calculation result for all matrix cells is obtained.

The results of the calculations can be seen in the last column of the last line of the matrix (the right-hand corner of the last line). From the calculation result, it is found that the distance of edit distance using Damerau-Levenshtein distance method from comparison of word "disertasl" and "dissertation" is 1. In table III.6 is shown comparison of edit distance resulting from all words recommended for word which is considered misspelled, ie the word "disertasl".

Table II. 5 Value Distance Distance For Disertasl Word Selection

Word Recommendation	Edit Distance
degradasi	6
dekoratif	7
dekorator	7
dekret	7
desertir	**4**
diakritik	6
disertasi	**1**
diskredit	5
disertai	**2**

Based on table III.6 there are three words whose edit distance is less than or equal to the maximum distance value. Of the three words will be selected which has the smallest edit distance value is the word "dissertation" with edit distance equal to one.

d. Mapping

The mapping process is comparing the original tokens with the fixed tokens of errors such as the process of altering characters based on Indonesian grammar rules, word-boundary errors, and non-word errors based on the index of both tokens.

The first stage mapping is done after the process of changing the characters based on the rules of Indonesian grammar as in the following figure.

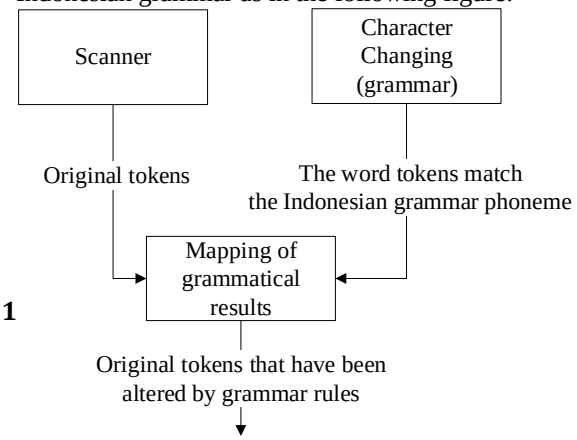


Figure II. 5 Mapping Results Changing Character Grammar Indonesia

The second stage mapping is done after the process of detection and correction of word-boundary spelling errors.

The third stage mapping is done after the word selection process for the correction of non-word spelling errors.

2.9 Testing System

The test is done with eleven dissertation documents, the following is the result of the whole test for the accuracy of detection and correction of each document word.

Table II. 6 Accuracy Detection Testing

File	Detection

	Recall	*Precision*
D 100 a.html	100%	20%
D 101 a.html	100%	50%
D 102 a.html	100%	25%
D 00901a.html	100%	54%
D 00903.html	100%	75%
D 00907.html	100%	50%
D 501.html	100%	76%
D 502.html	100%	76%
D 503.html	100%	80%
D 504.html	100%	78%
D 505.html	100%	71%
Average	100%	59,5%

Table II. 7 Accuracy Repair Testing

File	**Repair**	
	Recall	***Precision***
D 100 a.html	100%	100%
D 101 a.html	100%	50%
D 102 a.html	100%	100%
D 00901a.html	100%	100%
D 00903.html	100%	100%
D 00907.html	100%	100%
D 501.html	100%	38%
D 502.html	100%	53%
D 503.html	100%	63%
D 504.html	100%	10%
D 505.html	100%	71%
Average	100%	71,4%

3. CLOSING

3.1 Conclusion

Pengujian sistem yang dilakukan menunjukkan bahwa aplikasi yang yang dibuat untuk mendeteksi dan mengkoreksi kesalahan ejaan sudah memenuhi persyaratan fungsional. Dari tabel 4.8 analisis hasil pengujian deteksi kesalahan ejaan terdapat beberapa penyebab mengapa rata-rata nilai *precision* deteksi yang dihasilkan kecil sebesar 59,5%, yaitu:

1. Because of the name of the undetected person during the Name Entity Tagger process.
2. The words ends "nya", "kan".
3. Error in English words .

As for the correction test can be seen in table II.9 that the precision value for the correction is very small that is with an average of 71.4%. This is due to several causes ie:

1. The word recommended by soundex does not match the misspelled word. This is because if the first letter recommended by soundex is different from the misspelled word.
2. The replacement of the vowel character into consonant letters in the three-digit letter character will be encoded during the file conversion process .
3. Lack of Indonesian dictionary resources for words that have affixes at the beginning, middle and end of word .

3.2 suggestion

For further research, it is advisable to try other phonetic methods and use a better Named Entity Recoginition library for Indonesian to avoid being misspelled .

BIBLIOGRAPHY

[1] Amber Wilcox-O'Hearn, Graeme Hirst, and A. Budanitsky, "Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4919 LNCS, no. 2000, pp. 605–616, 2008.

[2] Mohammad Ali Elmi and Martha Evens, "Spelling correction using context," *Proc. ThirtySixth Annu. Meet. Assoc. Comput. Linguist. Seventeenth Int. Conf. Comput. Linguist.*, vol. 1, pp. 360–364, 1998.

[3] Behrang Qasemizadeh, Ali Ilkhani, and Amir Ganjeii, "Adaptive language independent spell cheeking using intelligent traverse on a tree," *2006 IEEE Conf. Cybern. Intell. Syst.*, 2006.

[4] Yinghao Huang and Yi Li Murphey, "Automotive diagnosis typo correction using domain Knowledge and Machine Learning," pp. 273–280, 2013.

[5] Utis Sutisna, "Koreksi Ejaan Query Bahasa Indonesia Menggunakan Algoritme Damerau Levenshtein," vol. 15, no. 2, pp. 25–29, 2010.

[6] Arumsari. K, "Penggunaan Metode Kesamaan String pada Pemeriksaan Ejaan Bahasa Indonesia," 1998.

[7] Fahmilu Kurniawan, "Implementasi Algoritme Fonetik Untuk Koreksi Ejaan Ejaan Bahasa Indonesia," 2010.

[8] Aminul Islam and Diana Inkpen, "Real-Word Spelling Correction using Google Web 1T 3-grams," *EMNLP '09 Proc. 2009 Conf. Empir. Methods Nat. Lang. Process. Vol. 3*, no. August, pp. 1241–1249, 2009.

[9] Andrew Carlson and Ian Fette, "Memory-based context-sensitive spelling correction at web scale," *Proc. - 6th Int. Conf. Mach. Learn. Appl. ICMLA 2007*, pp. 166–171, 2007.

[10] A. Insanudin, "Ekstraksi Informasi Kemacetan pada Media Digital," 2013.

[11] Abdul. Chaer, *Tata Bahasa Praktis Bahasa Indonesia*, Revisi, Ce. Jakarta: Rineka Cipta, 2011.

[12] Robert C. Russell, "Russell Soundex System,"

1918.
[13] Fred J. Damerau, “A technique for computer detection and correction of spelling errors,” *Commun. ACM*, vol. 7, no. 3, pp. 171–176, 1964.
[14] Ulrich Pfeifer and L. I. VI, “Searching Proper Names in Databases,” *October*, vol. 20, pp. 1–13, 1994.
[15] Fauzi, Rizki Hilmi ‘Pembangunan Kamus Bahasa Indonesia sebagai Sumber Daya Natural Language Processing (NLP) Bahasa Indonesia,’ vol. 2017, pp. 193–200.
[16] “Cambridge Dictionary.” [Online]. Available: https://dictionary.cambridge.org/. [Accessed: 20-Jan-2018].