COMP90049 Project 1: Waht kinda typoz do poeple mak?

I. INTRODUCTION

Provide suggested correction for a misspelled word is a thriving task. One approximate matching algorithm may work better for a specific type of errors, therefore, both finding the most efficient algorithms and defining spelling errors classifications are urged to developed simultaneously. The main aims of this research are to recognize specific patterns of typographical errors made by specific authors through some approximate matching algorithms.

II. SPELLING ERROR STRATEGIES

2.1 Spelling Errors Types

A spelling error is resulted from type the wrong sequence of letters to construct a word, neglecting the fact whether the author has prior knowledge for the correct spelling or not [1]. Typographical errors and orthographical errors are parts of spelling errors. Typographical errors or "typos" are made unintentionally by authors who know the correct spelling, while orthographical errors are the contrary [1].

Many types of typos have been introduced. In general, it depends on how many edits needed (single or multiple), in what kind of words (short or long), and in which character (first or Nth). The four basic typos are caused by, insertion, deletion, substitution, and transposition between letters [1].

2.2 Approximate Matching Algorithms

Matching algorithms are used to measure the similarities between words. This research applies three common algorithms, Edit Distance, N-Gram, and Soundex to detect typographical errors.

1. Edit Distance

An approach to count how many single edit (insertion, deletion, and substitution) needed to transform between two words [2]. In this research, Edit Distance refers to Levenshtein distance that uses parameter 0 for match and 1 for every edit, which 0 means perfect match. By this distance, the four basic typos could be detected.

2. N-Gram Similarity

This approach measures similarity based on the intersection of each N sub-sequence of words [3]. This research uses N-Gram similarity with N=2 that yields a proportion from 0 to 1 (perfect match). This method is considered effective to detect semantic errors.

3. Soundex

This approach transforms words into indexes based on sound similarity [3]. The best match then is determined through Edit Distance. This method is meant for phonetic matching.

III. METHODOLOGY

3.1 Datasets

There are two sets of misspelled words used in this research. The first one is from Wikipedia that contains typos from Wikipedia editors [4]. The second one is from Birkbeck corpus. It contains handwritten spelling errors, made by students, that read by a transcription machine [5]. Every set has its corresponding correct words list. The spelling correction strategy is applied by matching each record in the two lists with an English dictionary list that contains 370,099 words [6].

Total Number of	Datasets	
	Wikipedia	Birkbeck
Misspelled words	4,453	34,683
Unique misspell words	4,227 (94.9%)	32,670 (94.2%)
Unique correct words	4,124 (92.6%)	22,694 (65.4%)
Correct words not found	25	62
Misspell words found	322	5,661
Accuracy loss	383	5,685

Around 8.6% records in Wikipedia and 16.4% records in Birkbeck have either misspelled words found in dictionary or correct words not found in dictionary. These anomalies will contribute to the accuracy loss.

Misspell Vs Correct Length	Datasets			
	Wikipedia		Birkbeck	
Under length	1761	39,5%	16163	46,6%
1 letter	1640	93,1%	9802	60,6%
2 letters	111	6,3%	3823	23,7%
> 3 letters	10	0,6%	2538	15,7%
Equal length	1726	38,8%	11248	32,4%
Overlength	966	21,7%	7272	21,0%
1 letter	890	92,1%	5946	81,8%
2 letters	57	5,9%	1085	14,9%
> 3 letters	19	2,0%	241	3,3%

Misspelled words in both datasets mostly shorter than the intended ones. Birkbeck's dataset have a considerable proportion for three letters shortage. This research is focusing on errors in Wikipedia data, thus the following analysis will be done with this data until it is said differently.

3.2 Preliminary Analysis

The top three highest frequencies of errors in Wikipedia data is having the same unique characters but different words length. This indicates the existence of typos.

Unique Letters Differences	Length Comparison		
	Under length	Same length	Overlength
Less letters	517	217	9
1 character	11,4%	4,8%	0,2%
> 2 character	0,2%	0,1%	0,0%
Same letters	1219 27,4%	1281 28,8%	705 15,8%
More letters	25	228	252
1 character	0,6%	5,1%	5,2%
> 2 character	0,0%	0,0%	0,4%

The basic typos can be seen in each category. Insertion is mostly represented in 'overlength', deletion in 'under length', and substitution in 'same length'. The following table represents the first errors happened to misspelled words that have one excessive letter but same unique letters (overlength, same letters) that could reflect the insertion errors.

Typical Insertion	Total Words	Highest Frequency on	Misspelled Examples
repeated letter	294	repeat "l"	auxilliaries, fullfill
add vowel	219	after "l"/"r" or before "a"	natural, similiar
add "r"	29	after "e"	intergrated, perpertrated
add "c"	24	after "s"/"x" or before "e"/"t'	abscence, excecute
add "h"	20	after "c" or before "t"	architecht, charachter
add "s"	18	after "e"/"i" or before "t"/"c"	substract, liscense
add "n"	16	after "i"	binominal, indentical
others	46	add "t", "d", "l"	alledged, detatched

In this category, insertion errors are mostly because double press the same keys (repeated letter). While for deletion errors, it can be seen in following category ‘less length, less letters’ with one-letter difference.

Typical Deletion	Total words	Highest Frequency on	Misspelled Examples
“r”	60	after "t"	constuction, instuments
“e”	54	after "h"/"t"	approachs, infinit
“s”	38	after “n”/”i”/"e"	contruction, repectively
“h”	36	after "c'	caracterized, psycology
“a”	34	after "e"	endevors, sergent
“c”	34	after "a"	aquire, manufatured
others	215	missed "t", "l", "i",”u”,”n"	agriculure, commonweath

There is no specific pattern for deletion errors but some specific letters were missed more frequently than others. Similar condition is applied for substitution and transposition errors.

Substitute Direction	Misspelled Examples
“a” → “e”	acedemic
“f” → “v”	himselv
“i” → “e”	prestigeous
“e” → “a”	secratary
Transpose Direction	Misspelled Examples
“ns” → “sn”	agaisnt
“id” → “di”	consdiered
“rn” → “nr”	govenrment
“rn” → “nr”	tempaltes

This analysis provides early detection for the basic typos. Specific patterns of errors type can be distinguished through spelling correction strategies.

3.3 Spelling Correction Strategies

Spelling correction can be done by comparing each misspelled word with all records in dictionary then return the predictions based on the minimum distance or maximum similarity. This research runs the process in Python using package with the following base commands.

Methods	Python Calling Function
Edit Distance	editdistance.eval(string1, string2)
N-Gram	ngram.NGram.compare(string1, string2, N = 2)
Soundex	editdistance.eval(jellyfish.soundex(string1), jellyfish.soundex(string2))

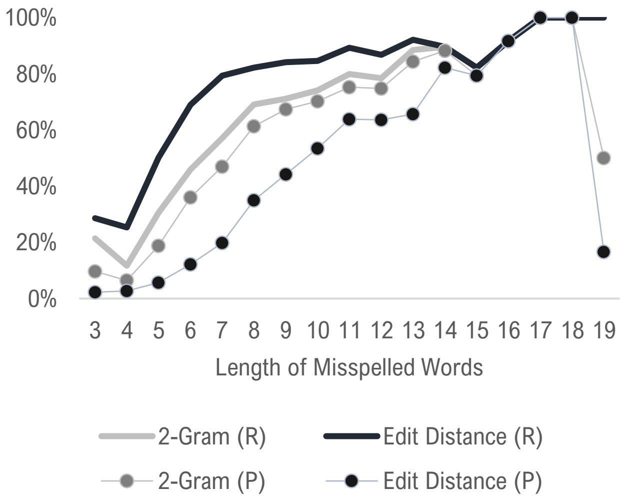
IV. RESULTS

4.1 Evaluation Metric

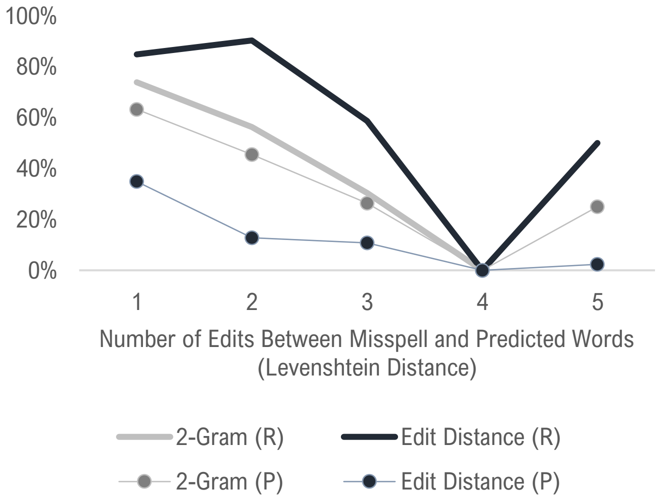
Accuracy rate shows the proportion of correct single predicted word in all records. It is highly dependent on the order of words in dictionary. Precision rate is the proportion of correct words in multiple predictions with the same best distance compare to total number of predictions. Recall rate has the same numerator as precision rate but use total misspelled records as the denominator.

Evaluation (Rate/Stats)	Edit Distance	2-Gram	Soundex + EditDistance
Accuracy	54.9% (2,444/4,453)	62.6% (2,787/4,453)	0.8% (34/4,453)
Precision	26.0% (3,520/13,516)	56.2% (2,916/5,192)	0.3% (3,606/1.2.10^6)
Recall	79.0% (3,520/4,453)	65.5% (2,916/4,453)	81.0% (3,606/4,453)
Average Distance	1	0.75	0
Maximum Distance	5	0.94	1
Average Predictions	3	1	272
Maximum Predictions	115	11	2,462

All the three methods return moderate to high recall but low precision. The highest precision rate is belong to N-Gram because it only returns few predicted words. Meanwhile, Soundex-based Edit Distance has the lowest one because it returns tremendous number of predicted words. Hence, this method is considered not effective for this dataset. The following graph shows Edit distance and 2-Gram performances based on the length of the misspelled words. R is for Recall and P is for Precision.



Both methods have lower precision in predicting small words (length ≤ 4) and increasing rate for longer words except for the longest one. The large gap between precision and recall for Edit Distance is because the correct word has higher probability to appear in the list but less precise. It represents a trade-off between precision and recall.



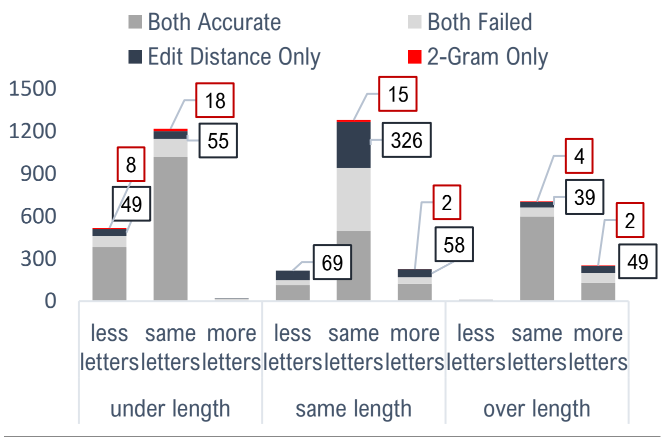
By comparing the frequency of edits between misspelled and predicted words and the evaluation rates, it can be inferred that most Wikipedia authors did basic errors. Both methods have their best performance for this type or errors and remain decreasing for multiple edit errors.

4.2 Performance Based on Typical Errors

The comparison of accuracy in the following discussion is based on the status whether the correct word exist in the predicted words or not exist.

Edit Distance	2-Gram		Total
	Failed	Accurate	
Failed	884	49	933
Accurate	653	2867	3520
Total	1537	2916	4453

According to the cross-tabulation, both methods are quite effective since they share around 65% accurate predictions. However, there are approximately 15% accurate predictions are solely supported by Edit Distance and around 1% by 2-Gram.



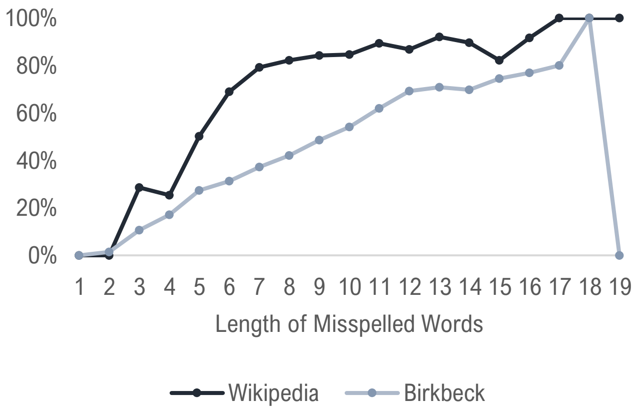
The success of both methods is distributed dominantly in category 'under length, same letters' and 'over length, same letters' which represent deletion and insertion errors. Edit Distance outperform in category 'same length, same letters' which represents substitution or transposition errors. This category and 'under length, same letters' which represents deletion errors are where the 2-Gram solo support mostly lies. These statistics strengthen the assumption that certain types of errors can only be predicted by certain algorithms.

misspelled words	Predictions	
	Edit Distance	2-Gram
auxilliaries	auxiliaries (√)	auxiliaries (√)
constuction	construction (√)	construction (√)
himselv	himself (√)	himself (√)
consdiered	considered (√) conspired	considered (√)
baceause	because (√)	basellaceae
conesencus	concentus consensus (√) noncensus	census
syphyllis	phyllis syphilis (√)	syllis
govement	movement	government (√)
honory	honor honora honors	honorary (√)
sophicated	sonicated spicated	sophisticated (√)

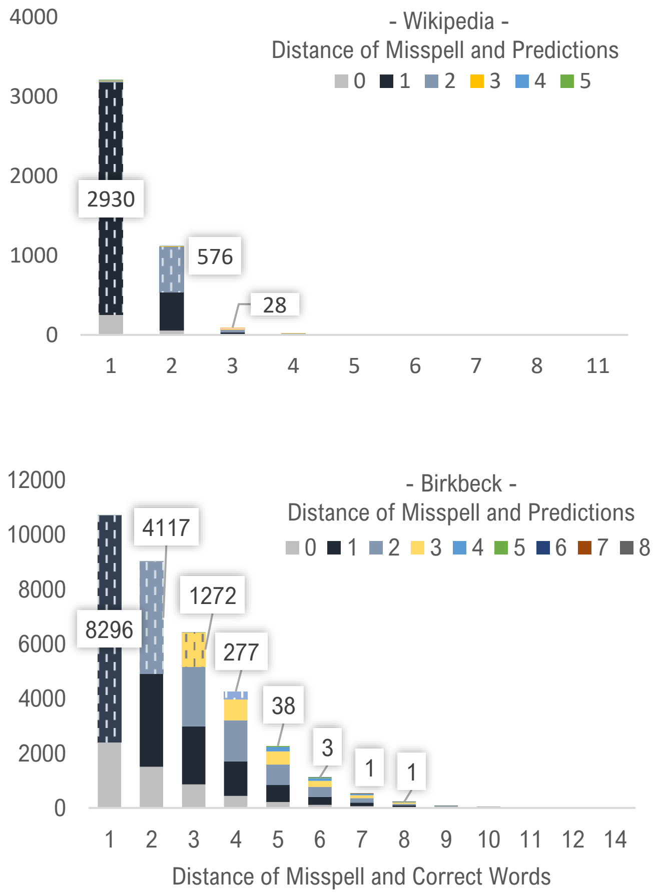
The above result answers the main research question to detect typical errors. In line with the previous analysis, there exist single edit errors and both methods work best for this type. There are also multiple edit errors. Edit Distance shows more accuracy for a combination of insertion and substitution errors (e.g. 'conesencus') than for the multiple deletion errors (e.g. 'govement'), in which 2-Gram tends to work better for these errors. The rest that are not correctly-predicted could be complex errors.

4.3 Performance Based on Typical Authors

Different characteristics of authors affect the performance of spelling correction. Recall of Edit Distance in Birkbeck dataset is merely 40.4%. Below is the distribution based on the length of the misspelled words compared to Wikipedia dataset.



The increasing trends for longer words are similar but recall is significantly lower than Wikipedia's. The following graph shows the comparison of the actual distance from misspelled into correct words and the predictive distance based on the prediction words.



The data labels highlight the accuracy where the actual distance same as the predictive distance. The accuracy loss because misspelled words found in dictionary is shown by predictive distance 0.

Wikipedia dataset is more representative for typographical errors. Since those are made unintentionally thus the actual distance merely small (1 – 3). Birkbeck dataset has more complex errors due to authors and machine limitations hence the actual distance lies within longer range (1 – 8).

The later graph clearly shows that most predicted words have smaller distance than it supposed to be and Edit Distance not well-performed for numerous edits. The following table are complex errors found in both datasets where the actual distance is greater than the predictive one.

Misspell Words	Correct Words	NE	D
Complicated Edits			
comcerracarlants	congratulations	11	B
suteranrepeiomene	subterranean	10	B
premonasterians	premonstratensians	6	W
Different words, Similar meaning			
macknifisent	splendid	10	B
mohammedans	muslims	8	W
unconfortability	discomfort	11	W
Different words, Different meaning			
ginderkarten	conscious	11	B
Same basic words, Different form			
seeked	sought	5	W
teached	taught	5	W
Excessive Under length			
efe	affectionately	12	B
charistics	characteristics	5	W

NE = 'Number of Edits' needed to transform misspell into correct word using Edit Distance
D = 'Dataset' B for Birkbeck and W for Wikipedia

V. CONCLUSION

Initial types of typographical errors are known through datasets exploratory and the detailed forms through approximate matching algorithms. Typographical errors can be in a form of single, multiple, and complex errors. Both Edit Distance and 2-Gram perform well to predict single edit errors for middle-length words, but each has its own unique capability. Edit Distance is more reliable for multiple substitution errors while 2-Gram for multiple deletion errors. Considering the number of predicted words, 2-Gram has more practical use than Edit Distance. Further analysis should be done by comparing more algorithms or by clustering.

REFERENCES

[1] Bhatti, Z., Ismaili, I. A., Shaikh, A. A., & Javaid, W. (2014). Spelling error trends and patterns in sindhi. *arXiv preprint arXiv:1403.4759.*

[2] Ahmed, B. (2015, July). Lexical normalisation of twitter data. In *Science and Information Conference (SAI)*, 2015 (pp. 326-328). IEEE.

[3] Buscaldi, D., Tournier, R., Aussenac-Gilles, N., & Mothe, J. (2012, June). Irit: Textual similarity combining conceptual similarity with an n-gram comparison method. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 552-556). Association for Computational Linguistics.

[4] Wikipedia contributors (n.d.) Wikipedia:Lists of common misspellings. In *Wikipedia, The Free Encyclopedia,* https://en.wikipedia.org/w/index.php?title=Wikipedia:Lists_of_common_misspellings&oldid=813410985

[5] Mitton, Roger (1980) Birkbeck spelling error corpus. In *University of Oxford Text Archive*,http://ota.ox.ac.uk/headers/0643.xml

[6] GitHub, Inc. https://github.com/dwyl/english-words