Automatic Extraction of Frequently Confused Words in English Based on String Similarity Algorithm

Weijie Kang*  
Heilongjiang University of Technology, China, 158100  
*Corresponding author e-mail: wqwangquan365@163.com

Abstract: The calculation method for the form similarity of English words is carried out. An algorithm where the upper limit of string similarity parameters can be set is used for automatic extraction of words with similar spellings from a specified vocabulary range. The frequently confused words with similar spellings screened can enrich the English lexical knowledge base after duplicate removal and classification. The frequently confused word knowledge base is of application value in the fields of textbook writing, vocabulary training design, dictionary compilation, and real word misspelling correction.

Keywords: Frequently Confused Words, Words with Similar Spellings, Lexical Knowledge Base, String Similarity, Form Similarity

1. Introduction
There are a lot of frequently confused words in English, where synonyms account for a significant proportion. For example, the nouns “hazard” and “peril” are both formal words, both of which mean “dangerous”, but peril is mostly used in literary works and refers to serious danger. Most synonymous nouns and verbs have no form similarity, but some conjugates [1-2], especially adjectives with the same root, have similar meanings and similar morphologies. For example, the adjectives “historical” and “historical” are both related to history, but “historical” means “historically important” or “historically famous”, and “historical” means “happened in history” or “historical” only. In addition to synonyms, there are many form similarities in English, but the words have very different meanings. For example, the adjective sensitive means “sensitive”, but sense means “sane.” Although sensitive and sense have common roots and the same parts of speech, these two words are not synonyms [3]. Laufer refers to words with similar forms as synforms. Words with similar spellings are not necessarily those with the same roots. Some words with similar spellings have the same part of speech, and some conjunctions have different parts of speech. For foreign language learners, words with similar spellings with only
one or two letters are very easy to confuse. For example, the adjective “ingenuous” means “simple”, while ingenious means “spiritual.” The homely of another pair of frequently confused adjectives is an adjective, the meaning of the word is “simple; ordinary”, while “homely” is a noun, meaning “long-course preaching”. In the teaching process, attention should be paid to guide learners to distinguish words with similar spellings\(^{[4]}\), to prevent the misuse of existing vocabulary knowledge for excessive generalization or false analogy. For example: the noun indifference is not the opposite of difference; the adjective inflammable is not the opposite of flammable; the past participle of crave is Craved, not craven (timid; timid). Even advanced learners need to consolidate and improve their vocabulary skills continually and avoid misinterpreting, misidentifying, misreading, or even using the incorrect words. If “prosecution” is misinterpreted as “persecution” during interpretation, it shall be corrected immediately. If the wrong words are used and not timely corrected, the results can be even worse. But how can we identify which words are frequently confused and adjectives that require attention?

Obviously, easy-to-blend adjectives based on experience accumulation and intuition are not only slow but also lack systematic approach. Hence, the calculation method of the form similarity between English words is researched here, and the quantitative methods are used to automatically extract the words with similar spellings from the vocabulary range of about four or eight levels of the English major (about 13,000 words). The frequently confused and words with similar spellings are stored in the thesaurus of frequently confused words to enrich the content of the English lexical knowledge base. The knowledge base of English promiscuous words has application value in the fields of textbook writing, vocabulary training design, dictionary compilation, and real-word spelling error checking.

2. String similarity algorithm

String similarity algorithm is a new tool for solving English learning problems with the help of optimization methods. It is especially suitable for the research of prediction and comprehensive evaluation problems. Its essence is a learning algorithm based on the optimization theory using a linear function hypothesis space in a high-dimensional feature space.

Assuming that it has linear separability, the sample set is \((x_i, y_i), i = 1, 2, \ldots, n\), \(x \in \mathbb{R}^d\), \(y \in \{-1, +1\}\). \(d\) the general form of a linear classification function in a dimensional space is \(g(x) = w \cdot x + b\), The specific classification surface equation is

\[
w \cdot x + b = 0
\]  

(1)

To ensure that the classification can be correctly classified in the face of all samples, the following equation shall be met

\[
y_i \left[ (w \cdot x) + b - 1 \right] \geq 0 \quad i = 1, 2, \cdots n
\]  

(2)

In summary, the minimum classification plane that satisfies the above conditions is the optimal
classification plane. The problem of solving the optimal classification surface by string similarity matching can be transformed into a constrained optimization problem, i.e., under the constraint condition of equation (2), and the minimum of the function can be obtained as follows

$$\varphi(w) = \frac{1}{2} \|w\|^2 = \frac{1}{2} (w \cdot w)$$

(3) Thus, the problem is further transformed into a Lagrange function as followed

$$L(w, b, a) = \frac{1}{2} (w \cdot w) - \sum_{i=1}^{n} a_i \left[ y_i (w \cdot x_i) + b \right] - 1.$$  

(4)

The optimal classification function obtained is as follows

$$f(x) = \text{sgn}\{ (w^* \cdot x + b^*) \} = \text{sgn}\left\{ \sum_{i=1}^{n} a_i^* y_i (x \cdot x_i) + b^* \right\},$$

(5) The sgn () is a symbolic function. In the case where the sample is strictly linearly separable, it can be processed according to the formulas (1) to (5); however, in the case where the sample is not linearly separable, a slack variable needs to be introduced, i.e., by adding a slack variable in formula (2) to meet the condition so that (2) can be used

$$y_i \left[ (w \cdot x) + b - 1 \right] + \xi_i \geq 0, i = 1, 2, \ldots, n ,$$

(6)

3. Automatic extraction and post-processing of words with similar spellings

The similarity between English words can be calculated by using the string similarity algorithm. However, the successful extraction of words with similar spellings also requires two thresholds to be determined according to the form characteristics of the promiscuous words with similar spellings: the length of the words to be queried and the string similarity function. The upper limit parameter in. Firstly, the words to be queried with less than 5 letters are excluded. As 4-letter and less than 4-letter words are less likely to be confused, and there are too many words that meet the conditions, the result of the extraction is not of great value for the induction of frequently confused words. During the extraction process, only the words with string similarity of 1, 2, and 3 are stored and classified, and all words with string similarity of 4 are ignored or excluded. The case where the string similarity is 0 is also not saved as the string similarity is 0 only when the spelling of two words is identical.

Although the process of extracting words with similar spellings is automatic, it still needs to be judged manually to filter out the frequently confused words. Displaying and saving the extraction results according to the similarity of strings is helpful for filtering. Take precede as an example. After classification, only the words with string similarity of 1 are obtained; there are 6 words with a distance of 2; and 48 words have a distance of 3 from “precede”, as shown in Figure 1. Hence, a group of words with similar spellings that require attention can be screened out: precede/proceed/precedent/accede/concede/cede/recede.
The more letters in the queried word, the higher the prominence of frequently confused words. Because the total number of matches is not large, it can be slightly deleted. For example, the word with string similarity of 1 to “ingenious” is only ingenious; a word with an ingenious distance of 2 is vacant; there are only 5 words with a distance of 3, as shown in Figure 2. The screening results are: ingenious/ ingenious/indigenous.

Not every word has a match with string similarity of 1. For example: eligible has no match with a distance of 1, as shown in Figure 3. However, the candidates with the similarity of 2 and 3 to the eligible string have the following words with similar spellings that may be confused: eligible/illegible/legible/ negligible.

In addition to being able to implement a single query function, it can also perform batch automatic extraction of words with similar spellings. The steps of batch extraction are similar to the steps of a single query, except that there is an additional loop processing link. FIG. 4 is a schematic diagram of an automatic batch extraction of words with similar spellings in a specified vocabulary.
The results of batch automatic extraction need to be manually filtered, deduplicated, and classified before they can be used as the content required for the frequently confused word classification database of the English lexical knowledge base. For example, eligible/illegible and illegible/eligible are actually frequently confused words with the same content; the screening results of reflective can be divided into three groups: reflective/irrespective/respect; reflective/respectful; prospective/perspective. Whether or not the latter two groups are retained should be determined as appropriate.

The frequently confused words with similar spellings can be further classified. The processing in the following sort order can mostly avoid crossovers.
4. Conclusions

String similarity is an essential basis for calculating and quantifying the form similarity of English vocabulary. Based on the similarity of strings, it is possible to extract frequently confused words from the syllabus vocabulary, the corpus vocabulary of the specified word frequency range, or other specified vocabulary ranges. The results of such extraction are systematic and quantitative, which are more reliable than mere empirical or intuitive induction. Through the study in this paper, only the automation of the extraction process is implemented, while manual processing and classification are still required in the subsequent processing stage. However, after the lexical knowledge base with classification annotation for frequently confused English words with similar spellings is established, the retrieval speed and efficiency will be significantly improved. Hence, the application prospect is broad.

References

[1] Xiang, W. L., Li, Y. Z., He, R. C., Gao, M. X., & An, M. Q. (2017). A novel artificial bee colony algorithm based on the cosine similarity. Computers & Industrial Engineering, 115, 54-68.
[2] Chen, Y., Liu, W., Xiong, Y., Duan, J., & Zhu, H. (2015). A fuzzy similarity elimination algorithm for indoor fingerprint positioning. International Journal of Distributed Sensor Networks, 2015, 1-10.
[3] Cao, B., Wang, J., Fan, J., Yin, J., & Dong, T. (2016). Querying similar process models based on the hungarian algorithm. IEEE Transactions on Services Computing, 10(1), 1-1.
[4] Wang, L., Xu, L., Yu, J., Xue, Y., & Zhang, G. (2016). Context-aware edge similarity segmentation algorithm of time series. Cluster Computing, 19(3), 1421-1436.