

Master Thesis: Useful Words

Leonard Paál

Hochschule Karlsruhe

Supervising professor: Jannik Strötgen

Additional supervision by: John Blake (University of Aizu)

December 2024

Abstract

Useful words are nice to have.

Contents

1	Evaluation of Approaches	3
1.1	Comparison of list similarities	3
1.1.1	Metrics used	3
1.1.2	Metrics rejected	4
2	Sources for existing lists for comparison	5

1 Evaluation of Approaches

1.1 Comparison of list similarities

In order to compare the results provided by the various approaches, metrics are needed that can be consistently calculated across the different approaches. While a human may be able to qualitatively analyze lists and gain a rough idea of their similarity, computed metrics provide an instantaneous (if simplified) outlook on similarities. A metric shall be defined as a function which takes as parameters two word lists of equal length which are word lists ordered descendingly by supposed relevance, and outputs a real number giving either a distance or similarity between the lists.

Considerations of the choice of metric

- **Metrics must be able to handle elements which occur in only one of the two lists.** Thus, a metric which solely compares ranks of elements is not viable.
- **Changes at the beginning of the lists have a greater impact on the metric output than changes further down.** The rationale is that two lists which have large differences at the start should be regarded as more different than ones where words
- **Changes in rank should be counted as more important if they occur further up the list.** A word that is rank 1 in list A but rank 101 in list B says more about the similarity than if a word is rank 2000 in list A but rank 2100 in list B.

1.1.1 Metrics used

Sequential rank agreement (modified) [1]: This metric is based on the deviations of some subset of the lists in the upper ranks. It is important to note that this metric has an additional parameter "depth" which determines how many elements (from the top of the list) are considered. It is therefore more helpful to view its results at various depths. The original formula for this metric in the case of two lists is:

$$a_d := a \text{ from start to rank } d$$

$$S_d := a_d \cup b_d$$

$$SRA_d(a, b) := \lambda \cdot \frac{\sum_{x \in S_d} \sigma^2((r_b(x)) - (r_a(x)))}{|S_d|}$$

where λ is a normalization factor ensuring that $\max(SRA) = 1$. In its proposed form, this metric can only compare lists which contain the same set of unique elements, just in different orders. In order to make it work on lists where this is not the case, one can set the "rank" of nonexisting elements to a value greater than the length of the lists, such as $2|a|$. Another drawback of the metric is that the standard deviation of two numbers does not depend on their absolute value, only their difference. However, to satisfy number 3 of the stated requirements, we can take the deviation of the logarithm of the ranks instead of the deviation of the ranks themselves, resulting in the formula

$$r'(x) := \begin{cases} \text{rank}_b(x) & \text{if } x \in b, \\ 2 \cdot |a| & \text{otherwise.} \end{cases}$$

$$SRA_d^{mod}(a, b) := \lambda \cdot \frac{\sum_{x \in S_d} \sigma^2(\log(r'_b(x)) - \log(r'_a(x)))}{|S_d|}$$

For this modified version, λ can be calculated with:

$$\lambda = \frac{1}{SRA_d(a, a^*)},$$

where a^* is a list such that $a \cap a^* = \emptyset$.

Discounted Cumulative Gain: This formula outputs a value between 0 and 1, with 1 being given if both lists are identical, 0 when they have no elements in common, and values in between when there is partial overlap between elements and/or their order is different. $DCG_p = \sum_{i=1}^p \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2(i+1)}$

$$rel_i := \begin{cases} \frac{1}{\text{rank}_b(el_i)+1} & \text{if } el_i \in b \\ 0 & \text{otherwise} \end{cases}$$

1.1.2 Metrics rejected

Kendall rank correlation [2]: This metric is bounded between 0 and 1 and compares the ranks of the elements of two lists. However, it cannot handle elements that only occur in one of the two lists, and thus is not suitable for our purposes. It also does not distinguish between differences in the upper and lower parts of the lists.

Spearman's footrule [3]: Rejected for the same reasons as Kendall rank correlation.

2 Sources for existing lists for comparison

It may be useful to compare the lists generated by the various approaches with existing word lists from educational materials: Textbooks often feature chapters with word lists, or sentences which can be converted to word lists with a tokenizer. The purpose is to have a point of comparison, to see if generated lists agree with existing lists, and find reasons for differences.

- Language learning textbooks
- Language learning applications
 - Duolingo: While Duolingo is the most popular language learning application as of 2024, it does not publish its word lists or course contents that is free of cost and easily convertible to a format that can be processed with NLP tools.
 - Rosetta Stone: Rosetta Stone publishes Course contents on its website. While the contents take the form of sentences, these can be converted to word lists by using a tokenizer on the contents and creating a list in the order in which they appear in the texts.

References

- [1] Claus Thorn Ekstrøm, Thomas Alexander Gerds, Andreas Kryger Jensen, and Kasper Brink-Jensen. Sequential rank agreement methods for comparison of ranked lists, August 2015.
- [2] M. G. KENDALL. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93, June 1938.
- [3] Charles Spearman. Correlation calculated from faulty data. *British journal of psychology*, 3(3):271, 1910.