Evaluating utility of vocabulary to language learners

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 Introduction **4**
- 1.1 Motivation . . . 5
  - 1.1.1 Role of vocabulary in language acquisition . . . 5
  - 1.1.2 Context-specific language learning . . . 5
  - 1.1.3 Examples of contexts and words . . . 6
- 1.2 Background . . . 8
  - 1.2.1 Requirements for calculating word utility from existing data . . . 8
  - 1.2.2 Developments in Natural Language Processing . . . 8
- 1.3 Method of this paper . . . 9
  - 1.3.1 Word extraction methods . . . 9

2 Literature Review **11**

3 Data **12**
- 3.1 Sources for existing lists for comparison . . . 13

4 Analysis **14**
- 4.1 Components of XAI word extraction . . . 15
  - 4.1.1 Formal problem statement for XAI word extraction . . . 15
- 4.2 XAI methods . . . 16
- 4.3 Tokenizers . . . 17
- 4.4 AI models . . . 18
- 4.5 Result evaluation & comparison . . . 19
  - 4.5.1 Comparison of list similarities . . . 19

5 Results **21**

6 Discussion **22**

7 Appendix **23**
- 7.1 Abbreviations . . . 24

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 Role of vocabulary in language acquisition

Learning a second language involves many different skills, often categorized into listening, reading, speaking, and writing. Another categorization may be vocabulary, grammatical skills, the ability to understand known words in various accents, understanding language when spoken at a fast speed. One skill that is required for any of these if the knowledge of vocabulary in the target language. A person with basic grammatical skills but no vocabulary has no ability to express themselves or understand anything which they hear around them. On the other hand, a person familiar with rudimentary vocabulary but no grammatical knowledge may struggle with understanding complex sentences and sound unnatural when speaking, but can at least make sense of short phrases and express themselves. Thus, basic knowledge of vocabulary is clearly one of the most essential skills for using a language. This raises the question of which vocabulary should be learned first when starting out on the journey of language acquisition.

1.1.2 Context-specific language learning

Learners of languages are typically interested in one or more specific aspects of the language. There is no such thing as an unbiased corpus than can fit every use case. Decisions must be made how much focus is given to everyday conversation, academic writing, writing pertaining to business like job applications etc. Many textbooks group vocabulary by topic, with a new topic being introduced with each lesson that student should ideally be able to converse in after completing the lesson. However, this way of introducing vocabulary has several shortcomings:

- Students spend time learning specific terms about the topic in one lesson while not learning even general vocabulary in other topics until much later.
- Knowing the words from previous lessons becomes a prerequisite for the more advanced material, especially because the terms from earlier lessons are used in example sentences for grammar. Thus, learners interested in learning the use of the language in one context will have a hard time of skipping earlier, less pertinent lessons to them.

Since the advent of computer-aided natural language processing, methods have been suggested to computationally identify useful words: Nation and Waring propose in an 1997 study [4] that the frequency with which a word appears in the target language should be used a metric for its importance to the learner, or utility, and thus teaching high-frequency words should be the focus when teaching beginners. However, focusing on maximum-frequency words to achieve text coverage (e.g., knowledge of 90% of words in a text) may not be as useful as one might at first think, as the most frequent words tend to be generic terms like “but”, “from”, “time”, “world” etc. Knowing only these words is not sufficient for comprehending most texts.

The TF-IDF metric [5] is often employed for finding the keywords in a document and thus a proxy for how important a word is for the overall meaning of the document. It essentially is a frequency normalized against a larger, background corpus and expressed whether the usage frequency in the document is unusual. Using TF-IDF

as a measure for utility of a word in a given context is possible, but it suppresses words that may be generally useful.

Thus, there is a need for further exploration as to how word utility can be calculated, using modern NLP methods involving artificial intelligence where necessary. Put more simply, this question may be reduced to:

What order or words, when learned, gives the learner the best set of words to understand and communicate in the language as quickly as possible?

1.1.3 Examples of contexts and words

Some examples for contexts that could be interesting to language learners include:

- Reading Wikipedia articles about a specific field (computer science, literature, biographies)
- Watching movies
- Travel to a country where the target language is spoken
- Doing business with a company from a country where the target language is spoken
- Cultural exploration (literature, religion)
- Finding friends from other countries

The different contexts for which learners might be motivated to learn a language differ in how easily corpora can be obtained about to mine patterns from. Movie subtitles and Wikipedia articles are easily obtained from sites such as opensubtitles.org and wikipedia.org. The words that might be relevant for travel are not as easily obtained: One might imagine an ideal scenario to collect data, in which a statistically relevant group of people travelling to the destination to be examined are randomly selected and equipped with microphones and cameras before the travel. During travel, one could record their conversations, conversations with people around them, and materials they attempt to read to navigate their journey such as train schedules, descriptions of tours, restaurant menus, street signs, etc. Lacking the funds to conduct an experiment for every possible language, this paper is interested in finding a methodology to obtain data from readily available corpora and websites online that extracts relevant vocabulary from the source texts. Depending on the context, we can think about which of the following English words might be likely to appear frequently in the texts:

- Convert
- Cash
- Hug
- Dammit
- Y'all

- From
- Nineen eighty four
- Married

To examine a few examples: Words like “convert” occur frequent when looking at Wikipedia articles [1]. “cash” is likely to be useful for travelers, but in most other contexts, it would not be as relevant. “Y’all” is almost never used in formal writing but used abundantly in everyday speech in the southern United Stated of America and South Africa. “From”, meanwhile, will be likely to be one of the most frequently used words regardless of context.

1.2 Background

1.2.1 Requirements for calculating word utility from existing data

Large word corpora in many languages

Hardware capabilities

Software is able to process text on semantic level Text processing software and tools have existed since the 1980s, however before the advent of neural networks and other artificial intelligence methods, their capabilities were mostly bereft of any semantic understanding of the text processed: Using exclusively manually crated tools, it is possible to create a program which recognizes that *table* and *tables* derive from the same word, but much harder to recognize that there is any connection whatsoever between *table* and *chair*. The language processing capabilities of any human dwarf those of these early tools.

1.2.2 Developments in Natural Language Processing

Sophisticated tokenizers

Wordnets

Neural networks trained in NLP tasks

Explainable AI

This paper aims to exploit these significant developments to gain insights into which words can provide second language learners with the most utility.

1.3 Method of this paper

It is evident that recent tools based on Artificial Intelligence are much better-equipped for many tasks in the realm of Natural Language Processing than rule-based tools. At the same time, it is less clear how exactly these networks achieve the results they do. Fortunately, there is a branch in the field of Artificial Intelligence called Explainable Artificial Intelligence which aims at explaining the outputs of AI models. Thus, it may be possible to harness the evident intelligence off AI models to simulate a human, to find out which words have the largest impact on the model's performance.

1.3.1 Word extraction methods

The approaches for approximating utility with an automatically computable metric which this works aims to compare include:

Traditional

Raw frequency of words in corpus A simple ordering of words by how often they appear in a corpus.

TF-IDF How often the words appear in a target corpus but divided by their frequency in a more generic corpus. This metric is typically used to employ the most relevant words in documents for identifying keywords that express best its core topic.

AI-simulated learner

Performance difference of AI for NLP tasks Here, a Large Language Model (LLM) or a more specific language processing model is made to run NLP tasks such as text summarization, sentiment detection or question-answering. To find out which words help the AI model the most in performing its tasks, words are methodically omitted from texts and the AI's performance is recorded. This metric attempts to approximate utility by finding words which, when missing, cause the greatest performance loss in the NLP tasks. Evaluation metrics like Shapley values[5] may be used to measure the impact of missing words

Transformer attention The transformer architecture is based on a mechanism called *self-attention*. It allocates the neural network's processing to important parts of the input and thus provides some degree of explainability "out of the box".

Difference in internal vector representation for AI reading text This approach words similarly to the above involving an AI model, but instead of measuring the changes in the quality of its output, it measures how much changing the input to the model changes its the internal vector state: AI stores data in vector format, and when performing NLP tasks on texts, there is an internal vector representation. By using various distance metrics, it may be possible to find out which words have the greatest impact on the model's understanding of a text. Most of these approaches can be done both for individual words and word sequences (n-grams). While individual words are the easiest to examine,

sometimes n-grams are insightful for finding sequences of words whose meaning is more than the sum of their parts (idioms and collocations) and which therefore must be learned in separately from their constituents (meaningful English n-grams include e.g. “kick the bucket”, “such that”, “such as”).

This also raises the question of what is considered a “word”. A phrase like “such as” can be considered two words if the definition of a word is simply “something separated by a space” or one word if the definition is “a phrase whose meaning cannot be arrived at trivially from knowing the definition of its parts”. In Natural Language Processing, tokenizers break down texts into words, but they typically use the first definition for a word in the case of English. Many non-European language do not use spaces in their spelling (e.g. Japanese, Mandarin Chinese) or use spaces to separate a different unit of text (syllables in Vietnamese, sentences in Thai), making this definition of a word unpractical. In most languages, words can appear in various different forms: Verbs in Spanish are conjugated according to the time and originator of an action, Nouns in German are declined depending on their number and grammatical case. This adds another variable for compiling word lists: Whether the list should consider any different combination of letters as a different word, or whether different forms of the same headword should be viewed as only one word.

Key technologies employed include therefore:

- Tokenizers
- Lemmatizers
- Translators
- AI models to perform NLP tasks

Chapter 2

Literature Review

Chapter 3

Data

3.1 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.

Chapter 4

Analysis

4.1 Components of XAI word extraction

Components are

- XAI method used
- Tokenizer
- (Pretrained) AI model used
- NLP task

It is readily seen that these components are not independent of each other. Some completely determine the choice of another, while others limit the selection of the other components. [describe dependencies between components] Task -¿ model -¿ tokenizer -¿ words. must investigate comparability of results later. (also corpus -¿ task) must investigate comparability of results.

4.1.1 Formal problem statement for XAI word extraction

Givens:

- A set W of w candidate words: $|W| = w$.
- A corpus C containing lines/sentences in the target language.
- A function f indicating the performance at the chosen task when given the subset of W

$$f : 2^W \to \mathbb{R}$$
$$f(K) \mapsto p$$

- An integer k denoting the desired cardinality of the (smaller) subset of words to learn.

Find

$$\arg\max_K f(K)$$

$$K \subset W$$
$$|K| = k$$
$$k < |W|$$

In practice, it is often not feasible to calculate calculate f for every possible subset K, necessitating the use of approximations.

4.2 XAI methods

- Attention as Explanation
- Single Token Ablation

4.3 Tokenizers

(intimately related to AI models)

4.4 AI models

- NSP-model ABC

4.5 Result evaluation & comparison

4.5.1 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 are:

Handling of lists with partial overlap. 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.

Start of lists is more impactful than end. Since the beginning of lists contains the words which are ranked as most important, changes at the top should impact the metric more than changes at the bottom. This includes (1) Equal differences 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. Likewise, if a word is absent from list B, it implies a greater difference if that word is at rank 1 in list A than if it were at rank 1000.

Metrics used

Sequential rank agreement (modified) [2]: 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:

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

$$S_d := a_d \cup b_d$$

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

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

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

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

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. $DCG_p = \sum_{i=1}^{p} \frac{rel_i}{\log_2(i+1)} = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2(i+1)}$

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

Metrics rejected

Kendall rank correlation [3]: 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 [6]: Rejected for the same reasons as Kendall rank correlation.

Chapter 5

Results

Chapter 6

Discussion

Chapter 7

Appendix

7.1 Abbreviations

NSP Next sentence prediction

NLP Natural language processing

LLM Large language model

Bibliography

[1] Leipzig Wortschatz: English Corpora. https://wortschatz.uni-leipzig.de/en/download/English.

[2] Claus Thorn Ekstrøm, Thomas Alexander Gerds, Andreas Kryger Jensen, and Kasper Brink-Jensen. Sequential rank agreement methods for comparison of ranked lists, August 2015.

[3] M. G. KENDALL. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93, June 1938.

[4] P. Nation. Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy/acquisition and pedagogy*, pages 6–19, 1997.

[5] Shahzad Qaiser and Ramsha Ali. Text mining: Use of TF-IDF to examine the relevance of words to documents. *International Journal of Computer Applications*, 181(1):25–29, 2018.

[6] Charles Spearman. Correlation calculated from faulty data. *British journal of psychology*, 3(3):271, 1910.