Association Measures

Jonathan Dunn
University of Illinois Urbana-Champaign

Key Points

- Association measures can find sequences of words with meaningful relationships
- Association measures and non-contextual embeddings are closely related representations
- But ranks of association are difficult to interpret without a larger quantitative framework

1. Introduction

Association measures are quantitative representations of the attraction or co-occurrence probability of two elements (henceforth, words). A high association means that two words occur together more often than we would expect by chance alone. This would imply, in turn, that there is some deeper linguistic relationship between these two words. A low association, on the other hand, would imply that two words are less likely to occur together than we would expect by chance, that they repel one another. This would imply instead that some linguistic relationship prevents those words from co-occurring. For example, the words "doctor" and "patient" are likely to have a high association because they occur within the same semantic frame. But "doctor" and "physician" are likely to have a low association because they are synonyms which would occupy the same slot within that frame. Thus, both high and low association values can tell us something about linguistic relationships.

This chapter starts with the basic standard for frequency-based association measures: pointwise mutual information (PMI). We will then consider how to deal with multi-unit association (Section 3), direction-specific association (Section 4), approximations of association through embeddings (Section 5), and how to interpret association values (Section 6).

Although this chapter focuses on the symmetrical PMI measure and the asymmetrical ΔP measure, there are many other types of statistical association measures in use. All these measures are attempts to operationalize the search for words which occur together beyond chance levels. For an in-depth statistical evaluation of different association measures, see Evert (2005). For a comparison of association measures for the purpose of extracting collocations, see Pecina (2009). The goal of this present chapter is to outline two representative types of association measure and then compare them with related work on word embeddings from computational linguistics, representations which can often be used for the same tasks.

2. Basic Co-Occurrence

The basic idea of Pointwise Mutual Information (PMI: Church and Hanks, 1990) is to compare the probability of two words occurring together with the joint probability of their occurring independently. In the simplest version, *co-occurrence* means that two words are adjacent to one another. However, we could easily define a window size so that words that occur near each other are also counted. The equation below shows the basic calculation for PMI: the probability that two words (x and y) occur together over the joint probability of each in the corpus. This latter probability is the chance baseline: how often would these words occur together by accident?

$$\text{(1a) } pmi(x,y) = \log \frac{p(x,y)}{p(x)*p(y)}$$

Given this definition of PMI, we can improve it with a few adjustments. First, most applications use the Positive Pointwise Mutual Information (PPMI). This is because negative values are supposed to represent pairs of words that repel one another, that occur together less often than random chance would predict. Second, the range of PMI values will vary by corpus. This is because larger corpora have more words and more word sequences; thus, larger corpora have to spread the same amount of probability over more pairs, leading to smaller probabilities as an artifact simply of the size of the corpus. A Normalized PMI (NPMI: Bouma, 2009) corrects for this so that all values fall between 0 (no association) and 1 (high association). Notice, also, that the PMI works with log probabilities; this is because the probabilities involved are generally quite small.

This chapter provides an example based on 1 million sentences from news sources.[1] Table 1 below shows some of the most associated and least associated pairs of words in this corpus. The left column shows the most associated words without a frequency threshold and the right with a frequency threshold of 25. The contrast shows why frequency also needs to be accounted for: the most associated pairs on the left are combinations of very rare words. Because these rare words never occur with other words, they are highly associated. But that does not mean that they are particularly interesting pairs. This table shows three layers of association: the highest pairs in this particular case are named entities (top, around 0.99); the middle pairs are collocations (center, around 0.49), and the lowest pairs are parts of grammatical constructions (bottom, around 0.21). At all three levels, the pairs on the right are much more meaningful than the pairs on the left. This shows the practical importance of using a frequency threshold together with the NPMI.

Table 1. NPMI With and Without Frequency Thresholds

Min Frequency = 1		*Min Frequency = 25*	
abafazi wathint imbokodo	1.00	buenos aires	0.99
renier meintjies	1.00	recep Tayyip	0.99
bethereum currency	1.00	kuala lumpur	0.99
hosin mabeti	1.00	mardi gras	0.99
magali debatte	1.00	burkina faso	0.99
adverse effect	0.49	pick up	0.49
teller machine	0.49	army corps	0.49
year-old girl	0.49	border closures	0.49
annual passholders	0.49	vastly different	0.49
annual snowbird	0.49	hospitality sector	0.49
soft emotional	0.21	act like	0.21
soft corner	0.21	which could	0.21
rocket israel	0.21	are pending	0.21

[1] The code and corpus data are available here: https://doi.org/10.17605/OSF.IO/98XJ7

fridays article	0.21	much effort	0.21
operational problems	0.21	was assaulted	0.21

3. Multi-Unit Association

The default approach is to measure association between two words. But this is an arbitrary limitation and we would miss many interesting patterns if we simply looked at pairs of words. We have several approaches to generalize across longer sequences (c.f., Dunn, 2018), and here we will investigate two such methods. The first is based on *connector words*. These are function words like "to" or "for" or "the" which are so common that they are unlikely to be particularly associated with any given phrase. The simplest approach is to simply ignore these functional connector words and add them for free to any adjacent pair. Table 2 below shows the kinds of phrases we would get using connector words to create longer sequences of association. Some of these are phrases (like "tip of the iceberg") while others are named entities (like "guardians of the galaxy").

Table 2. NPMI with Connector Words

Phrase	***NPMI***
tip of the iceberg	0.77
filing with the securities	0.73
dividend on an annualized	0.70
stock in a transaction	0.66
guardians of the galaxy	0.66
bold and the beautiful	0.63
disclosure with the securities	0.62
completion of the transaction	0.62
young and the restless	0.61
possession of a firearm	0.61

A second approach is to recursively process the corpus, at each step joining together associated sequences and then treating these as atomic units. Thus, in the first pass we might learn that "pick up" is one collocation and "dirty laundry" is another. In the second pass, the same NPMI measure is calculated, but now viewing "pick_up" and "dirty_laundry" as individual words. The output of this algorithm would be increasingly long sequences of associated words, such as "pick up dirty laundry". Thus, this approach is recursive because it repeats the same underlying algorithm on the output of a previous iteration. For the examples shown in Table 3, we have computed this for five cycles; with each cycle we decrease the frequency and association thresholds used to identify two words as a single unit. The phrases shown in the table are largely named entities. This shows how the parameters involved shape the output (the frequency and association threshold required to join words into a single unit). The problem of multi-unit association has been explored in more detail in Dunn (2018). In general, higher frequencies reflect grammatical rather than world knowledge patterns; and higher association values reflect world knowledge rather than grammatical patterns. The third important component here is context: in which linguistic or non-linguistic settings does a particular pattern occur? Thus, the fuller picture would combine information about association strength with information about frequency, both within the scope of specific contexts. In this case, the context is news articles and we would expect different sets of phrases for the same speakers when observed using language in different contexts.

Table 3. Multi-Unit Association

Merged Phrases	*NPMI*
nissay asset management corp	0.98
sen mitt romney rutah	0.91
all progressives congress apc	0.87
centers for disease control and preventions	0.83
secretary of state mike pompeo	0.77
teetering on the brink	0.73
respond to a request for comment	0.72
democratic presidential candidate former vice president joe biden	0.68
orange county registrar of voters	0.68
did not immediately respond	0.66

4. Asymmetrical Association

Attraction is not symmetrical: for a pair of words like “of course”, the probability of seeing “course” given “of” is quite low (there are many prepositional phrases starting with “of”). But the probability of seeing “of” first given “course” is quite high (“course” is relatively rare and the phrase “of course” accounts for much of its usage). The PMI, however, provides a single measure regardless of direction. An alternative measure is the ΔP which instead provides two direction-specific measures (Ellis, 2007).

Table 4. Quantities for Calculating Directional Association

	Y Present (Y_P)	*Y Absent (Y_A)*	*Totals*
X Present (X_P)	a	b	a + b
X Absent (X_A)	c	d	c + d
Totals	a + c	b + d	

The variables necessary to measure direction-specific association are shown in Table 4. Imagine we have two words, “of” and “course”. The row *X Present* contains all the counts with “of” and the row *X Absent* contains all the counts without “of”. This leads to frequency-based variables that measure how many times each permutation occurs in the corpus. These are combined into direction-specific associations in the equations in (2a) and (2b) below. For instance, the left-to-right measure compares the frequency of “of course” with the total frequency of “of”: in how many cases is “of” followed by “course”? A lower probability will in turn mean that other continuations are more likely. The resulting measures are comparable to the PMI, but focusing on either left-to-right or right-to-left continuations.

$$\text{(1a) } \Delta P_{LR} = \frac{a}{(a+c)} - \frac{b}{(b+d)}$$

$$\text{(1b) } \Delta P_{RL} = \frac{a}{(a+b)} - \frac{c}{(c+d)}$$

The advantage of the ΔP is that we can distinguish between two sub-cases of association: cases in which the first word is dependent on the second and cases in which the second word is dependent on the first. This is illustrated in Table 5 below. The examples on the left have a high LR association but a very low RL association and the examples of the right show the reverse, with a high RL association. For example, phrasal verbs (with their very frequent particles/prepositions) have a low LR association but are captured with the RL measure. These different behaviors show the advantage of being able to split association by direction when necessary. For a fuller analysis of the relationship between association measures and specific behaviors like sentence processing, see Wiechmann (2008).

Table 5. Examples of Asymmetric Association

LR Dominant			*RL Dominant*		
new zealanders	0.996	0.001	pursuant to	0.000	0.971
at gunpoint	0.993	0.000	ramped up	0.002	0.973
per annum	0.998	0.009	afoul of	0.000	0.972
their newly-acquired	0.992	0.003	cooped up	0.001	0.973
new yorkers	0.987	0.004	stave off	0.003	0.977
life expectancy	0.979	0.004	epitome of	0.000	0.974
white supremacists	0.986	0.012	coincided with	0.000	0.974
in accordance	0.970	0.001	plethora of	0.000	0.974
per capita	0.987	0.026	irrespective of	0.000	0.975
the slightest	0.941	0.000	inclement weather	0.019	0.997

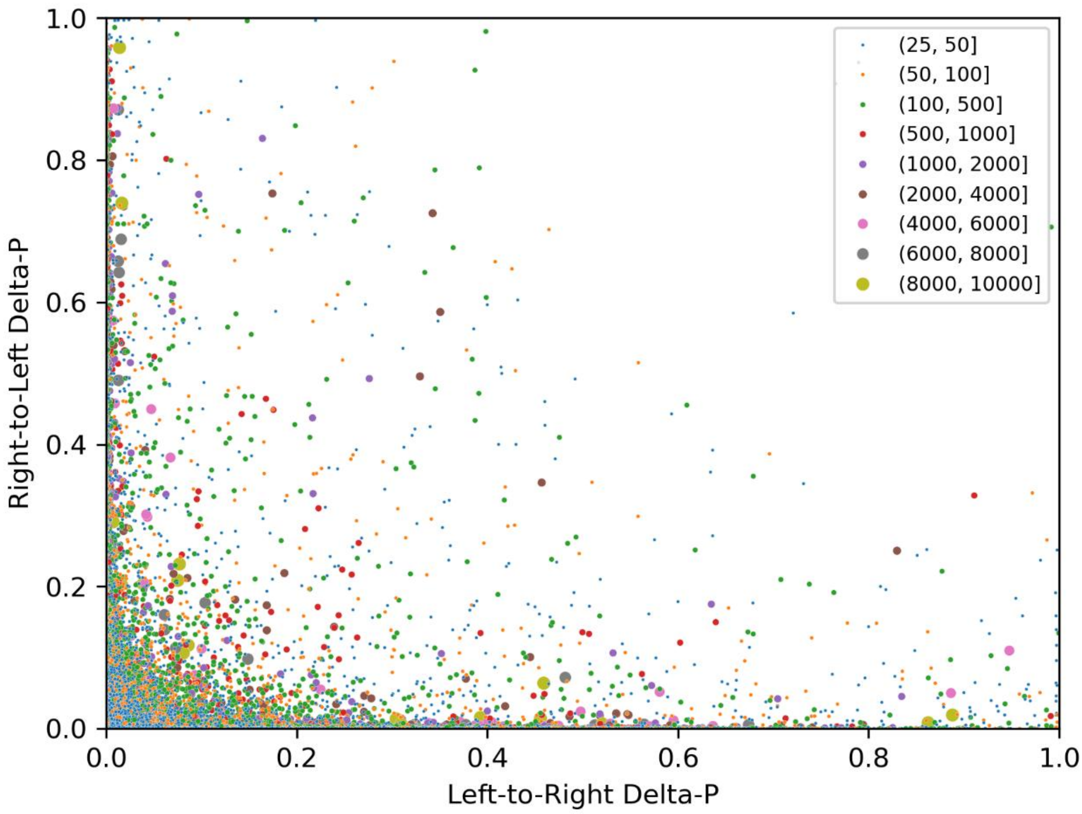

Figure 1. Scatterplot of Differences in Direction-Specific Association

While Table 5 shows only a few examples, the relationship between left-to-right and right-to-left association together with frequency is shown in Figure 1. The y-axis shows right-to-left association, with values at the top being highly associated; and the x-axis shows left-to-right association, with values to the right being highly associated. Frequency strata are represented by color as well as point size: larger points are more frequent. Each point is a phrase, like “per capita”. Most points are clustered in the bottom left of the figure, meaning that they have low association. In general, only a small percentage of phrases show high association in both directions, although many such phrases do exist. This figure thus provides a more detailed view of the relationship between directional association and frequency. Measures like the PMI would collapse the y-axis and x-axis here into a single dimension.

5. Association and Embeddings

While we have previously focused on frequency-based association measures, this section introduces and compares model-based embeddings. Both types of representations aim to capture underlying linguistic relationships between words. Association measures capture co-occurrence between pairs of words; thus, we could create a co-occurrence matrix with each word as a row and as a column. This kind of matrix would represent the overall association between each word in the vocabulary. The challenge is that, for even a moderately sized corpus, this matrix would be very large. This can be contrasted with model-based embeddings, which capture the same information as pairwise association measures but in a compressed format.

Why? Imagine a corpus with a substantial vocabulary of 10,000 words: the co-occurrence matrix would contain 10k rows and 10k columns, for a total of 100 million cells. If the association measure is symmetrical, we could reduce this by half to 50 million word pairs. A matrix of this size scales poorly. Thus, a long-standing approach is to use dimension reduction to convert this distance matrix into embedding representations for each word (GloVE: Pennington, et al., 2014). For this reason, there is a close theoretical connection between association measures and embedding representations (Levy, et al., 2015).

Here we compare NPMI as a frequency-based measure of association with skip-gram and continuous-bag-of-word embeddings (trained with negative sampling: Mikolov, et al., 2013). These are approximations of the kinds of embeddings derived from dimension reduction and are more commonly used because they can be estimated from larger corpora. Skip-gram embeddings (commonly abbreviated SG) are approximations based on the idea that a word should be able to predict what other words occur in its context. And continuous-bag-of-word embeddings (CBOW) are approximations based on the idea that a word’s context should be able to predict it.

Our basic approach is to get NPMI and cosine similarities between words from the same corpus; cosine similarities are distances between embedding representations, where we can think of the embeddings as reductions of a co-occurrence matrix. To the degree that these two representations capture the same information, the ranks of pairs of words should be similar. In other words, if “pick up” has a high association, these words should also be located near one another in the embedding space. The question here is whether the same information is captured in both types of representations, with embeddings being more efficient at scale.

Table 6. NPMI vs Cosine Similarity in Skip-Gram Embeddings

Words	*NPMI*	*Cosine*
mardi gras	0.99	0.92
burkina faso	0.99	0.95
los angeles	0.99	-0.23

chiang mai	0.82	0.81
mick jagger	0.82	0.81
san diego	0.82	-0.02
charles dickens	0.65	0.59
hurricane katrina	0.65	0.50
rabbit hole	0.65	0.51
white supremacists	0.65	0.09
san francisco-based	0.65	-0.06
fatal accident	0.44	0.55
denver colorado	0.44	0.68
infrastructure projects	0.44	0.62
camera sensor	0.44	0.78

We test this relationship using both skip-grams (window size = 5) and continuous-bag-of-words (cbow: window size = 2). The Pearson correlation between NPMI and cbow embeddings is a highly significant 0.61; with skip-gram embeddings it is a highly significant 0.62. We explore some examples in Table 6. Each series of examples shows those which align across the two measures as well as some outliers. The complete results show that the correlation is generally high when outliers are removed.[2] For example, "los angeles" is highly associated but distant in the embedding space; this is true for other place names like "san diego" as well. At the bottom, we see examples that are closer in the embedding space than with association measures, cases like "fatal accident". These examples show how we get a slightly different view based on whether we use frequency-based association or model-based embeddings.

What factors lead to a non-perfect correlation between association measures and similarity within an embedding space? An important factor is that these embeddings are less stable because they are estimated rather than based on observed frequencies (Hellrich, et al., 2019). For instance, the negative sampling algorithm chooses random words as negative comparisons for each positive instance of co-occurrence. This randomness leads to less stable representations, especially on smaller corpora like the 1 million sentences we are using in this example. Overall, however, the two methods lead to similar results.

Given that association measures and word embeddings largely agree in the ranking of sequences of words, what is the importance of embeddings? First, embeddings are an essential component of modern computational linguistics and it is important to see that these are closely related to association measures. Second, embeddings provide a more convenient format for storing and querying co-occurrence information. For example "fatal accident" has relatively similar association in both approaches. However, with embeddings it is easy to expand that sequence to other similar sequences like "deadly accident", simply by finding phrases that are nearby within the embedding space.

6. Interpreting Association

Measuring the association between words, whether through frequency-based measures or through embeddings, is a way to observe patterns in language. However, we are ultimately interested in the factors which cause a high or low association, not with the measures

[2] https://doi.org/10.17605/OSF.IO/98XJ7

themselves. What are some of the possible reasons for a high association and what are the confounding factors? We illustrate the challenge of interpretation in Table 7 by showing similar phrases by NPMI: one adpositional phrase and one verb phrase.

Table 7. Ranks of Association Within Similar Phrases

“around the” + Noun		*“take” + X*	
around the world	0.51	take advantage	0.53
around the globe	0.50	take place	0.49
around the clock	0.45	take care	0.40
around the corner	0.41	take precautions	0.38
around the country	0.31	take action	0.36
around the nation	0.19	take responsibility	0.35
around the same	0.18	take effect	0.34
around the region	0.17	take their	0.12
around the league	0.17	take months	0.11
around the area	0.16	take this	0.11
around the house	0.16	take every	0.11
around the city	0.16	take out	0.10
around the state	0.12	take my	0.10

This table shows us, for instance, that “around the world” and “around the globe” are highly associated while “around the city” and “around the state” have a much lower association. This reflects idiomatic usage, in that a phrase like “around the world” has more than its literal meaning and is used to refer to wide-ranging events or phenomena. On the right side of the table, we see similar ranks of verb phrases with “take”. The top examples are idioms like “take advantage” or “take care” and the bottom examples, with quite low association, are neither idioms nor syntactic bundles but rather almost random co-occurrences like “take my”.

The point here is that association measures can help us to sort such phrases, but we cannot use them to distinguish between causes of high association like idioms or syntactic units. And it is difficult to set a fixed threshold, above which co-occurrences are interesting and below which they are not. For these reasons it is important that association as a means of measuring linguistic relationships be used within a larger quantitative model with specific hypotheses to be tested, rather than as a means of exploration without a larger framework. In other words, association measures in isolation do not provide an exhaustive analysis.

7. Summary

Association measures like Pointwise Mutual Information (PMI) allow us to find sequences of words which have meaningful linguistic relationships. These measures can include more than two words, can observe non-adjacent co-occurrence, and can be direction-specific. These are three important attributes for improving our view of patterns in a corpus. From both a theoretical and an empirical perspective, association measures are similar to non-contextual embeddings like skip-grams. The main interpretive challenge with these representations is to untangle the different causes of high association and find out what they mean linguistically.

References

Bouma, G. (2009). “Normalized (pointwise) mutual information in collocation extraction.” In Christian Chiarcos and Richard Eckart de Castilho and Manfred Stede, ed., *Proceedings of the German Society for Computational Linguistics and Language Technology*. Gunter Narr Verlag, 31-40.

Church, K. and Hanks, P. (1990). “Word association norms, mutual information, and lexicography.” *Computational Linguistics,* 16(1), 22-29.

Dunn, J. (2018). “Multi-unit association measures: Moving beyond pairs of words.” *International Journal of Corpus Linguistics,* 23(2), 183-215.

Ellis, N. (2007). “Language Acquisition as Rational Contingency Learning.” *Applied Linguistics* 27(1), 1-24.

Evert, S. (2005). *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. Ph.D. Thesis. Stuttgart: University of Stuttgart.

Hellrich, J., Kampe, B. and Hahn, U. (2019). “The Influence of Down-Sampling Strategies on SVD Word Embedding Stability.” In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. Association for Computational Linguistics, 18-26.

Levy, O., Goldberg, Y. and Dagan, I. (2015). “Improving Distributional Similarity with Lessons Learned from Word Embeddings.” *Transactions of the Association for Computational Linguistics,* 3: 211-225.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. and Dean, J. (2013). “Distributed Representations of Words and Phrases and Their Compositionality.” In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, *Volume 2*. Curran Associates Inc. 3111-3119.

Pecina, P. (2009). “Lexical Association Measures and Collocation Extraction.” *Language Resources and Evaluation*, 44(1/2): 137-158.

Pennington, J., Socher, R. and Manning, C. D. (2014). “GloVe: Global Vectors for Word Representation.” In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 1532-1543.

Wiechmann, D. (2008). “On the Computation of Collostructional Strength: Testing Measures of Association as Expressions of Lexical Bias.” *Corpus Linguistics and Linguistic Theory*, 4(2): 253-290.