Two Multivariate Generalizations of Pointwise Mutual Information

Tim Van de Cruys
RCEAL
University of Cambridge
United Kingdom
tv234@cam.ac.uk

Abstract

Since its introduction into the NLP community, pointwise mutual information has proven to be a useful association measure in numerous natural language processing applications such as collocation extraction and word space models. In its original form, it is restricted to the analysis of two-way co-occurrences. NLP problems, however, need not be restricted to two-way co-occurrences; often, a particular problem can be more naturally tackled when formulated as a multi-way problem. In this paper, we explore two multivariate generalizations of pointwise mutual information, and explore their usefulness and nature in the extraction of *subject verb object* triples.

1 Introduction

Mutual information (Shannon and Weaver, 1949) is a measure of mutual dependence between two random variables. The measure – and more specifically its instantiation for specific outcomes called pointwise mutual information (PMI) – has proven to be a useful association measure in numerous natural language processing applications. Since its introduction into the NLP community (Church and Hanks, 1990), it has been used in order to tackle or improve upon several NLP problems, including collocation extraction (ibid.) and word space models (Pantel and Lin, 2002). In its original form, it is restricted to the analysis of two-way co-occurrences. NLP problems, however, need not be restricted to two-way co-occurrences; often, a particular problem can be more naturally tackled when formulated as a multi-way problem. Notably, the framework of tensor decomposition, that has recently permeated into the NLP community (Turney, 2007; Baroni and Lenci, 2010; Giesbrecht, 2010; Van de Cruys, 2010), analyzes language issues as multi-way co-occurrences. Up till now, little attention has been devoted to the weighting of such multi-way co-occurrences (which, for the research cited above, results either in using no weighting at all, or in applying an ad-hoc weighting solution without any theoretical underpinnings).

In this paper, we explore two possible generalizations of pointwise mutual information for multi-way co-occurrences from a theoretical point of view. In section 2, we discuss some relevant related work, mainly in the field of information theory. In section 3 the two generalizations of PMI are laid out in more detail, based on their global multivariate counterparts. Section 4 then discusses some applications in the light of NLP, while section 5 concludes and hints at some directions for future research.

2 Previous work

Research into the generalization of mutual information was pioneered in two seminal papers. The first one to explore the interaction of multiple random variables in the scope of information theory was McGill (1954). McGill described a first generalization of mutual information based on the notion of conditional entropy. This first generalization, called *interaction information*, is described in section 3.2.1 below. A second generalization, solely based on the commonalities of the random variables, was described by Watanabe (1960). This generalization,

Proceedings of the Workshop on Distributional Semantics and Compositionality (DiSCo'2011), pages 16–20,

called *total correlation* is presented in section 3.2.2.

3 Theory

3.1 Mutual information

Mutual information is a measure of the amount of information that one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to the knowledge of the other.

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \quad (1)$$

Pointwise mutual information is a measure of association that looks at particular instances of the two random variables X and Y. More specifically, pointwise mutual information measures the difference between the probability of their co-occurrence given their joint distribution and the probability of their co-occurrence given the marginal distributions of X and Y (thus assuming the two random variables are independent).

$$pmi(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \quad (2)$$

Note that mutual information (equation 1) yields the expected PMI value over all possible instances of random variables X and Y.

$$\mathbb{E}_{p(X,Y)}[pmi(X,Y)] \quad (3)$$

Furthermore, note that PMI may be positive or negative, but its expected outcome over all events (i.e. the global mutual information) is always non-negative.

3.2 Multivariate mutual information

In this section, the two generalizations for multivariate distributions are presented. For both generalizations, we examine their standard form (which looks at the interaction between the random variables as a whole) and their specific instantiation (that looks at particular outcomes of the random variables). Analogously to PMI, it is these specific instantiations of the measures that are able to weigh specific co-occurrences according to their importance in the corpus. As with PMI, the value for the global case ought to be the expected value for all the instantiations of the specific measure.

3.2.1 Interaction information

Interaction information (McGill, 1954) – also called co-information (Bell, 2003) – is based on the notion of conditional mutual information. Conditional mutual information is the mutual information of two random variables conditioned on a third one.

$$\begin{aligned} & I(X;Y|Z) \\ = & \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \end{aligned} \quad (4)$$

which can be rewritten as

$$\sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x,y,z) \log \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \quad (5)$$

For the case of three variables, the interaction information is then defined as the conditional mutual information subtracted by the standard mutual information.

$$\begin{aligned} I_1(X;Y;Z) &= I(X;Y|Z) - I(X;Y) \\ &= I(X;Z|Y) - I(X;Z) \\ &= I(Y;Z|X) - I(Y;Z) \end{aligned} \quad (6)$$

Expanded, this gives the following equation:

$$\begin{aligned} & I_1(X;Y;Z) \\ = & \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x,y,z) \log \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} \\ & - \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \quad (7)$$

We can now define *specific interaction information* as follows[1]:

[1]Note that – compared to equation 7 – the two subparts in the right-hand side of the equation have been swapped. For the three-variable case, this gives exactly the same outcome except for a change in sign. The swap is necessary in order to ensure a proper set-theoretic measure (Fano, 1961; Reza, 1994).

$$SI_1(x,y,z) = \log \frac{p(x,y)}{p(x)p(y)} - \log \frac{p(z)p(x,y,z)}{p(x,z)p(y,z)} = \log \frac{p(x,y)p(y,z)p(x,z)}{p(x)p(y)p(z)p(x,y,z)} \quad (8)$$

Interaction information – as well as specific interaction information – can equally be defined for $n > 3$ variables.

3.2.2 Total correlation

Total correlation (Watanabe, 1960) – also called multi-information (Studený and Vejnarová, 1998) quantifies the amount of information that is shared among the different random variables, and thus expresses how related a particular group of random variables are.

$$I_2(X_1, X_2, \ldots, X_n) = \sum_{\substack{x_1 \in X_1, \\ x_2 \in X_2, \\ \cdots \\ x_n \in X_n}} p(x_1, x_2, \ldots, x_n) \log \frac{p(x_1, x_2, \ldots, x_n)}{\Pi_{i=1}^{n} p(x_i)} \quad (9)$$

Analogously to the definition of pointwise mutual information, we can straightforwardly define the correlation for specific instances of the random variables, which we coin *specific correlation*.

$$SI_2(x_1, x_2, \ldots, x_n) = \log \frac{p(x_1, x_2, \ldots, x_n)}{\Pi_{i=1}^{n} p(x_i)} \quad (10)$$

For the case of three variables, this gives the following equation:

$$SI_2(x,y,z) = \log \frac{p(x,y,z)}{p(x)p(y)p(z)} \quad (11)$$

Note that this measure has been used in NLP tasks before, notably for collocation extraction (Villada Moirón, 2005).

4 Application

In this section, we explore the performance of the measures defined above in an NLP context, viz. the extraction of salient *subject verb object* triples. This research has been carried out for Dutch. The Twente Nieuws Corpus (Ordelman, 2002), a 500M Dutch word corpus, has been automatically parsed with the Dutch dependency parser ALPINO (van Noord, 2006), and all *subject verb object* triples with frequency $f \geq 3$ have been extracted. Next, a tensor $\mathcal{T}$ of size $I \times J \times K$ has been constructed, containing the three-way co-occurrence frequencies of the I most frequent subjects by the J most frequent verbs by the K most frequent objects, with $I = 10000, J = 1000, K = 10000$. Finally, two new tensors $\mathcal{U}$ and $\mathcal{V}$ have been constructed, such that $\mathcal{U}_{ijk} = SI_1(T_{ijk})$ and $\mathcal{V}_{ijk} = SI_2(T_{ijk})$, i.e. tensor $\mathcal{U}$ has been weighted using specific interaction information (equation 8) and tensor $\mathcal{V}$ has been weighted using specific correlation (equation 11).

Table 1 shows the top five *subject verb object* triples that received the highest specific interaction information score, while table 2 gives the top five *subject verb object* triples that gained the highest specific correlation score (both with $f > 30$).

Note that both methods are able to extract salient *subject verb object* triples, such as prototypical *svo* combinations (*peiling geeft opinie weer* 'poll represents opinion', *helikopter vuurt raket af* 'helicopter fires rocket') and fixed expressions (Dutch proverbs such as *de wal keert het schip* 'the circumstances change the course' and *de vlag dekt de lading* 'the content corresponds to the title').

subject	verb	object	SI_1
peiling 'poll'	*geef weer* 'represent'	*opinie* 'opinion'	18.20
helikopter 'helicopter'	*vuur af* 'fire'	*raket* 'rocket'	17.57
Man 'man'	*bijt* 'bite'	*hond* 'dog'	17.15
verwijt 'reproach'	*snijd* 'cut'	*hout* 'wood'	17.10
wal 'quay'	*keer* 'turn'	*schip* 'ship'	17.01

Table 1: Top five *subject verb object* triples with highest *specific interaction information* score

Comparing both methods, the results seem to indicate that the extracted triples are similar for both weightings. This, however, is not consistently the case: the results can differ significantly for partic-

subject	verb	object	SI_2
verwijt	*snijd*	*hout*	8.05
'reproach'	'cut'	'wood'	
helikopter	*vuur af*	*raket*	7.75
'helicopter'	'fire'	'rocket'	
peiling	*geef weer*	*opinie*	7.64
'poll'	'represent'	'opinion'	
vlag	*dek*	*lading*	7.21
'flag'	'cover'	'load'	
argument	*snijd*	*hout*	7.17
'argument'	'cut'	'wood'	

Table 2: Top five *subject verb object* triples with highest *specific correlation* score

subject	verb	object	SI_2
nationaliteit	*speel*	*rol*	4.12
'nationality'	'play'	'role'	
afkomst	*speel*	*rol*	4.06
'descent'	'play'	'role'	
toeval	*speel*	*rol*	4.04
'coincidence'	'play'	'role'	
motief	*speel*	*rol*	4.04
'motive'	'play'	'role'	
afstand	*speel*	*rol*	4.02
'distance'	'play'	'role'	

Table 4: Top five combinations with highest *specific correlation* scores for verb *speel*

ular instances. This becomes apparent when comparing table 3 and table 4, which for each method contain the top five combinations for the Dutch verb *speel* 'play'.

Table 3 indicates that specific interaction information picks up on prototypical *svo* combinations (*orkest speelt symfonie* 'orchestra plays symphony'; also note the 4 other triples that come from bridge game descriptions). Specific correlation (table 4), on the other hand, picks up on the expression *een rol spelen* 'play a role', and extracts salient subjects that go with the expression.

subject	verb	object	SI_1
orkest	*speel*	*symfonie*	11.65
'orchestra'	'play'	'symphony'	
leider	*speel*	*ruiten*	10.29
'leader'	'play'	'diamonds'	
leider	*speel*	*harten*	10.20
'leader'	'play'	'hearts'	
leider	*speel*	*schoppen*	10.01
'leader'	'play'	'spades'	
leider	*speel*	*klaveren*	9.89
'leader'	'play'	'clubs'	

Table 3: Top five combinations with highest *specific interaction information* scores for verb *speel*

In order to quantitatively assess the aptness of the two methods for the extraction of salient *svo* triples, we performed a small-scale manual evaluation of the 100 triples that scored the highest for each measure. A triple is considered salient when it is made up of a fixed (multi-word) expression, or when it consists of a fixed expression combined with a salient subject or object (e.g. *argument snijd hout* 'argument cut wood'). The bare frequency tensor (without any weighting) was used as a baseline. The results are presented in table 5.

measure	precision
baseline	.00
SI_1	.24
SI_2	.31

Table 5: Manual evaluation results for the extraction of salient *svo* triples

The results indicate that both measures are able to extract a significant number of salient triples compared to the frequency baseline, which is not able to extract any salient triples at all. Comparing both measures, *specific correlation* clearly performs best (.31 versus .24 for *specific interaction information*).

Additionally, we computed Kendall's τ_b to compare the rankings yielded by the two different methods (over all triples). The correlation between both rankings is $\tau_b = 0.21$, indicating that the results yielded by both methods – though correlated – differ to a significant extent.

These are, of course, preliminary results, and a more thorough evaluation is necessary to confirm the tendencies that emerge.

5 Conclusion

In this paper, we presented two multivariate generalizations of mutual information, as well as their instantiated counterparts *specific interaction information* and *specific correlation*, that are useful for weighting multi-way co-occurrences in NLP tasks. The main goal of this paper is to show that there is not just one straightforward generalization of pointwise mutual information for the multivariate case, and NLP researchers that want to exploit multi-way co-occurrences in an information-theoretic framework should take this fact into account.

Moreover, we have applied the two different measures to the extraction of *subject verb object* triples, and demonstrated that the results may differ significantly. It goes without saying that these are just exploratory and rudimentary observations; more research into the exact nature of both generalizations and their repercussions for NLP – as well as a proper quantitative evaluation – are imperative.

This brings us to some avenues for future work. More research needs to be carried with regard to the exact nature of the dependencies that both measures capture. Preliminary results show that they extract different information, but it is not clear what the exact nature of that information is. Secondly, we want to carry out a proper quantitative evaluation on different multi-way co-occurrence (factorization) tasks, in order to indicate which measure works best, and which measure might be more suitable for a particular task.

Acknowledgements

A number of anonymous reviewers provided fruitful remarks and comments on an earlier draft of this paper, from which the current version has significantly benefited.

References

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):1–48.

Anthony J. Bell. 2003. The co-information lattice. In *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA 2003*.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information & lexicography. *Computational Linguistics*, 16(1):22–29.

Robert Fano. 1961. *Transmission of information*. MIT Press, Cambridge, MA.

Eugenie Giesbrecht. 2010. Towards a matrix-based distributional model of meaning. In *Proceedings of the NAACL HLT 2010 Student Research Workshop*, pages 23–28. Association for Computational Linguistics.

William J. McGill. 1954. Multivariate information transmission. *Psychometrika*, 19(2):97–116.

R.J.F. Ordelman. 2002. Twente Nieuws Corpus (TwNC), August. Parlevink Language Techonology Group. University of Twente.

Patrick Pantel and Dekang Lin. 2002. Discovering word senses from text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada.

Fazlollah M. Reza. 1994. *An introduction to information theory*. Dover Publications.

Claude Shannon and Warren Weaver. 1949. *The mathematical theory of communication*. University of Illinois Press, Urbana, Illinois.

M. Studený and J. Vejnarová. 1998. The multiinformation function as a tool for measuring stochastic dependence. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pages 261–297, Norwell, MA, USA. Kluwer Academic Publishers.

Peter D. Turney. 2007. Empirical evaluation of four tensor decomposition algorithms. Technical Report ERB-1152, National Research Council, Institute for Information Technology.

Tim Van de Cruys. 2010. A non-negative tensor factorization model for selectional preference induction. *Natural Language Engineering*, 16(4):417–437.

Gertjan van Noord. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42, Leuven.

Begoña Villada Moirón. 2005. *Data-driven identification of fixed expressions and their modifiability*. Ph.D. thesis, University of Groningen, The Netherlands.

Satosi Watanabe. 1960. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4:66–82.