Statistical Measures for Characterising MWEs

Ismail El Maarouf, and Michael Oakes

i.el-maarouf@wlv.ac.uk, michael.oakes@wlv.ac.uk

Introduction

- Automatic identification of MWE
- New text distance-based measures.
- Comparison with standard association measures

Research context

- Corpus Pattern Analysis (Hanks, 2013), DVC project.
- The Pattern Dictionary of English Verbs (http://pdev.org.uk)
- Identification, representation and annotation of MWEs

Measures of word association and flexibility for the extraction of MWEs

1. Measuring strength of collocations with **Pointwise Mutual Information**

$$PMI(x,y) = log_2 \frac{P(x,y)}{P(x).P(y)}$$

Where P(x,y) is the probability of two words occurring in a common context (e.g. span of 5 words, or in subject-object relation), and P(x) and P(y) are the probabilities of finding words x and y respectively anywhere in the corpus.
PMI is positive if the two words tend to co-occur, 0 if they occur together as often as one would expect by chance, and less than 0 if they are in complementary distribution (Church and Hanks, 1989).

2. Two widely used association measures

T-score

$$Tscore(x,y) = \frac{F(x,y) - \frac{F_x.F_y}{N}}{\sqrt{F_{xy}}}$$

logDice

$$logDice(x,y) = 14 + log_2 D = 14 + log_2 \frac{2F_{xy}}{F_x + F_y}$$

3. Measuring flexibility of collocations using **Shannon's Diversity Index (Entropy)**

Mean μ of text distances

$$\mu_{(X,Y)} = \frac{1}{n}\sum_{i=1}^{n} dist(X_i, Y_i)$$

Standard Deviation σ of text distances

$$\sigma_{(X,Y)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(dist(X_i,Y_i) - \mu_{(X,Y)})^2}$$

Entropy E of text distances

$$E_{(X,Y)} = -\sum_{i=1}^{n} P_j log_2 P_j$$

4. Measuring **Idiomaticity** of collocations

$$Idiomaticity_{(X,Y)} = \frac{number\ of\ idiomatic\ occurrences\ of\ (X,Y)}{total\ number\ of\ occurrences\ of\ (X,Y)}$$

Statistical scores for 10 MWEs

Idiom	Freq	PMI	T-score	Log-Dice	Idiomaticity	Entropy	Mean length	Standard deviation
[back][bite]	87	5.914	10.38	5.549	0.989	0.338	1.057	0.277
[bullet][bite]	36	10.484	6.477	8.561	1	1.069	2.055	0.404
[head][bite][off]	30	6.009	7.639	5.6	0.775	3.281	3.032	2.721
[bug][bite]	19	10.589	4.688	7.894	0.842	3.326	3.125	2.578
[hand][bite][BENEFIT]	15	5.584	7.639	5.196	1	2.463	5.933	5.842
[bean][spill]	40	10.947	6.705	8.917	0.952	0.37	2.025	0.987
[straw][grasp/clutch]	33	9.865	6.077	8.172	0.892	2.213	3.485	1.623
[way][wind][blows]	21	10.663	25.264	10.652	0.676	2.488	3.5	0.534
[shoe/boot][quake/shiver/shake]	12	5.043	5.056	5.608	1	2.057	3.417	1.382
[bucket][kick]	5	8.647	4.349	7.004	0.238	0.721	3.5	1.87

Comparison using Pearson's correlation

	Freq	PMI	t-score	Log-Dice	Idiomaticity	Entropy	Mean length	Standard deviation
Freq	1							
PMI	-0.13	1						
t-score	0.1	0.2	1					
Log-Dice	-0.13	0.92 *	0.53	1				
idiomaticity	0.47	-0.26	-0.09	-0.22	1			
Entropy	-0.31	-0.37	-0.04	-0.32	0.12	1		
Mean length	-0.73*	-0.25	0.01	-0.2	-0.13	0.55	1	
Standard deviation	-0.46	-0.41	-0.28	-0.51	-0.03	0.51	0.84*	1

Conclusions

- Most statistically significant correlation (p = 0.000185, cor = 0.92) was between PMI and logDice.
- Significant correlation between the mean and the standard deviation of the length in words (p = 0.0022, cor = 0.84)
- Inverse correlation between frequency and mean length (p = 0.0174, cor = -0.73)
- New measures may not be useful for discovering new MWEs, but useful for characterising MWEs.
- Future work: (1) Confidence limits, (2) Apply to other languages

References

- Kenneth W. Church and Patrick Hanks. 1989. *Word Association Norms, Mutual Information and Lexicography.* In Proc. ACL: 76-83.
- Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations.* MIT Press.
- Michael P. Oakes. 2012. *Describing a Translational Corpus.* In Oakes, M. P. and Ji, M., *Quantitative Methods in Corpus-Based Translation Studies.* John Benjamins: 115-148.
- Pavel Pecina. 2008. *Lexical Association Measures: Collocation Extraction.* PhD thesis, Charles University in Prague.
- Pavel Rychlý. 2008. A lexicographer-friendly association score. In Proc. of RASLAN 2008: 6-9.

This work was partly supported by a AHRC grant [DVC, AH/J005940/1, 2012-2015]

PARSEME

University Alexandru Ioan Cuza, Iași, Romania