Automatic Extraction of MWEs for the Pattern Dictionary of English Verbs

Patrick Hanks, Ismail El Maarouf, and Michael Oakes

patrick.w.hanks@gmail.com, i.el-maarouf@wlv.ac.uk, michael.oakes@wlv.ac.uk

Introduction

- Automatic identification of MWE
- Word association Measures (Pecina, 2008)
- Idioms in the British National Corpus

Research context

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

Measures of word association and flexibility

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. Extending word association measures to 3 variables (Van de Cruys, 2011)

Specific Correlation for three variables

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

Specific Interaction Information for three variables

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

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

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

Case study: Statistics of word association and flexibility for *bite*

PMI in BNC50

verb x	collocate y	MI(x,y)
bitten	bug	12.49
biting	bullet	11.42
bite		11.25
bites		11.12
bitten		8.82
bit		7.46
bite	feeds	8.46
bites	off	5.38
bitten		4.66
biting		4.36
bite		3.92
bit		0.13

SC and SII in BNC50 (SII sorted)

y	z	SC	SII	Freq.
dobj:bullet	prep:over/variety	19.05	4.99	2
dobj:bullet	nsubj:maker	14.60	0.55	2
nsubj:maker	prep:over/variety	17.37	0.23	2
advmod:hard	prep:by/attacker	14.53	-0.29	2
dobj:bullet	partmod:take	13.28	-0.77	2
nsubj:recession	prep:into/industry	13.61	-1.08	2
partmod:take	prep:over/variety	16.05	-1.10	2
dobj:tongue	prt:off	7.80	-1.25	2
advcl:chew	prt:off	11.90	-1.46	4
auxpass:be	cc:and	-9.16	-2.47	2
advmod:even	aux:can	-1.96	-10.11	2
...	...	...	...	...

Flexibility of idioms in BNC50

X,Y	Freq.	μ	σ	E
bite, bullet	9	3	0	0
bite, feed*	5	4	0	0
bite, off*	6	4	0	1.057
bitten, bug	4	3.75	1.5	2

* including variants

Idiomaticity in the full BNC

X,Y	Idiom.	Freq.	Idiomaticity
bite, bullet	31	31	1
bite, bug	16	25	0.64
kick, bucket	6	15	0.4
spill, beans	40	41	0.98

PDEV entry for *bite*: 22 patterns, 10 idioms

13 *Pattern:* Idiom. **Human 1 bites** Human 2's **head off**
Implicature: Human 1 speaks sharply and unkindly to Human 2
Example: Just to ***bite*** their heads off.

14 *Pattern:* Idiom. **Human bites lip**
Implicature: Human grips his or her lip firmly with the teeth +
Example: He ***bit*** his lip but stood his ground.

15 *Pattern:* Idiom. **Human bite off more than [[Human]] can chew**
Implicature: Human undertakes a task that is is too difficult for him or her to accomplish successfully
Example: By aiming to depict Life in the 1990s, Kasdan has probably ***bitten*** off more than he can chew, but he

16 *Pattern:* Idiom. **Human bites** the **hand that feeds [[Human]]**
Implicature: Human attacks his or her benefactor +
Example: It is hard to ***bite*** the hand that feeds you.

17 *Pattern:* Idiom. **Human | Institution bites the bullet**
Implicature: Human *or* Institution decides to do something necessary but unpleasant +
Example: So, this week, Priddle ***bit*** the bullet.

18 *Pattern:* Idiom. **Human is bitten by the [MOD] bug**
Implicature: Human becomes very interested in [MOD]
Example: Chubby, bubbly jazzman Fats Waller was among the first to really get ***bitten*** by the London bug.

19 *Pattern:* Idiom. **Human bites the dust**
Implicature: Informal. Human dies suddenly and violently
Example: They ***bite*** the dust with lead in their bellies.

20 *Pattern:* Idiom. **Entity** *or* **Process bites the dust**
Implicature: Informal. Entity *or* Process comes to a sudden and unwelcome end
Example: If so, then we must freely admit that another time-honoured tradition of British self-restraint has very r

21 *Pattern:* Idiom. **Human bites** REFLDET **tongue**
Implicature: Human makes a desperate effort not to say what is in his or her mind
Example: It's all very well telling someone to ***bite*** their tongue and not fight back.

22 *Pattern:* Idiom. **once bitten twice shy**
Implicature: an unpleasant experience causes someone to be more cautious in future
Example: This time around it is a case of `once ***bitten***, twice shy' and their doubt is not simple but compound.

Perspectives

- Continue experiments on the use of statistical measures for MWEs.
- Combine measures in a statistical classifier for MWE extraction.
- Experiment with other languages (less fixed word order)

References and Acknowledgements

This work is partially supported by AHRC [DVC, AH/J005940/1, 2012-2015].

- **Pattern Dictionary of English Verbs:** `http://pdev.org.uk`
- Kenneth W. Church and Patrick Hanks. 1989. *Word Association Norms, Mutual Information and Lexicography.* 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.
- Tim Van de Cruys. 2011. *Two Multivariate Generalizations of Pointwise Mutual Information.* Disco 2011, 24 June 2011, Portland, OR.