Type-based Idiom Extraction

Jan Bušta

based on

A Measure of Syntactic Flexibility for Automatically Identifying Multiword Expressions in Corpora — by Colin Bannard

and

Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations — by Afsaneh Fazly and Suzanne Stevenson

25. 5. 2010, UdS

Out line

1) competition

2) object of interests

3) main goal

4) fixed ness

5) results I

6) mutual information

7) results II

8) conclusion

9) evaluation of competition

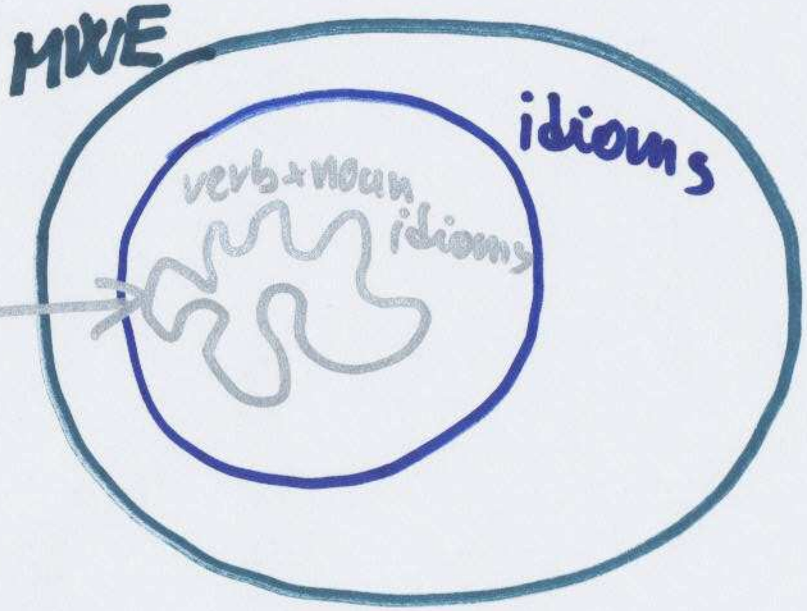
Objects of interests

Idiomatic phrases in form of

verb + noun

- transitive verb (needs an object)
- noun (with determiner

We will deal only with this type (verb+noun) in this presentation



Main goal

Is the phrase an idiom?

- determining fixedness
  -> level of idiomacity

- computing mutual information

Fixedness

(non) morphological variation of the phrase

- determiner

  run the show → run their show

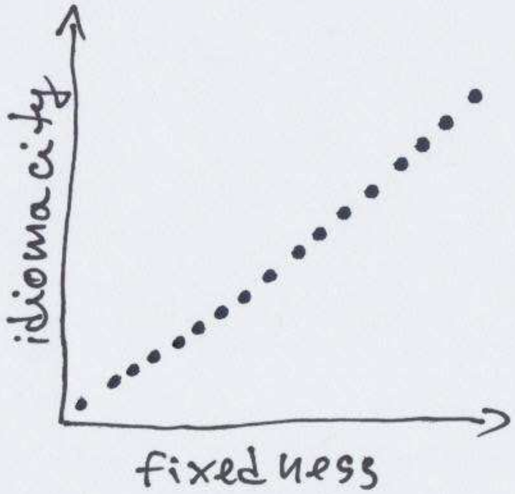
- internal modification

  break the ice → break the diplomatic ice

- passivisation

  call the shots → the shots were called by ...

- pluralisation



Fixedness II

- lexical

$$PMI(v, n_j) = \log \frac{P(v, n_j)}{P(r)\cdot P(n_j)}$$

$$F_{lex}^{(v,w)} = \frac{PMI(v,w) - \overline{PMI}}{s}$$

the mean

the standard deviation

<u>n_j</u> is a synonym for <u>w</u>

- syntactic

probabitity of pattern <u>pt</u> given <u>v</u> and <u>w</u>

$$F_{syn}^{(v,w)} = D(\overbrace{P(pt \mid v,w)} \,\|\, \underbrace{P(pt)})$$

probabitity of pattern <u>pt</u>

Kullback Lieber divergence

- hybrid

$$F_{hyb}(v,w) = \alpha F_{syn}(v,w) + (1-\alpha) F_{lex}(v,w)$$

preference of F_{syn} and F_{lex}

Results I

Measure	Accuracy	Relative error reduction
Random	50%	-
PMI	64%	28%
F_{lex}	65%	30%
F_{syn}	70%	40%
F_{hyb}	74%	48%

- canonical form
  - -> by set up threshold border
  - -> determining from patterns set

$$Cf_k(v,w) = \frac{f(v,w,pt_k) - \bar{f}}{s}$$

$pt_k \in$ set of patterns for the phrase $\underline{v}, \underline{w}$

Mutual information

- amount of information in bits that $\underline{y}$ provides about $\underline{x}$ given $\underline{z}$ (and vice versa)

$$I(x;y\,|z) = \underset{\text{box}}{\overset{\text{black}}{\blacksquare}} = \log_2 \frac{p(x\,|\,y,z)}{p(x\,|\,z)}$$

- syntactic variation

$$\mathrm{SynVar}(W) = \sum_i I(\mathrm{VerbVar}_i \,;\, \mathrm{Obj} \,|\, \mathrm{Verb}) + \sum_j I(\mathrm{ObjVar}_j \,;\, \mathrm{Verb} \,|\, \mathrm{Obj})$$

Resalts II

- using mutual information (with frequency) is better than $\underbrace{\text{t-score, MI-score, ...}}_{\text{they are based on frequency}}$

- combining determiner variation, internal modification and passivisation goes to best results than frequency based scores, but combining freq. with P, I, D is the best

Conclusion

Advantages

- both techniques are robust
- they work independent on dictionary

Disadvantages

- bigger corpus => better results
- verb + noun phrases ONLY!

Evaluation of competition

- play the second violin
- sleep on laurel
- have a chicken brain
- have a handle/have a window
- have big eyes
- finish with his own hand
- ask for hand
- draw the same rope
- spit to the water well
- pull at hair
- go to the dog-rose
- have a juice
- be behind the water
- have hare's plans
- admit of color

And the winner is...

Any questions?

:-)

Thank you for your attention.

xbusta@fi.muni.cz