Leveraging distributed representations and lexico-syntactic fixedness for token-level prediction of the idiomaticity of English verb–noun combinations

Milton King and Paul Cook

University of New Brunswick
Fredericton, Canada

Multiword Expressions

- Expressions of multiple words that can exhibit an idiomatic meaning
  - *Ivory tower*
  - *Hit up*
  - *Take a walk*
- Verb noun combinations
  - *See stars*
  - *Kick the bucket*

Idiomatic vs Literal

- Pull plug
  - (I) *They pulled the plug on the Department of Health funding*
  - (L) *Unfortunately someone pulled the sink plug*
- See stars
  - *(I) It caught him on the head and he went down seeing little sparkling stars*
  - *(L) It's still dark enough to see the brightest stars*

Idiom Token Classification

- Determine if an MWE instance is idiomatic
  - *They pulled the plug on the project* ➡ *[Idiomatic/Literal]*
- Applications
  - Machine translation
    - *Kick the bucket* ➡ *[mourir/frapper avec le pied]*
  - Sentence completion
    - *Keegan is ready to pull the plug on [a deal / the tv]*

Overview of Approach

- Supervised approach
- VNC token instances are represented via use of an embedding model
- Embedding models
  - Skip-thoughts
  - Word2vec
  - Siamese CBOW
- SVM classifier

Lexico-Syntactic Fixedness

- The idiomatic meaning of an expression is typically restricted to a small number of lexico-syntactic patterns
- **See star** (Idiomatic)
  - Active voice, no determiner, plural noun
    - *See stars*
- **See star** (Literal)
  - Active voice, determiner, singular noun
    - *See a star*
  - Passive voice, plural noun
    - *Stars were seen*

Patterns

Pattern No.	Pattern Signature			Example
1	v_{act}	det:NULL	n_{sg}	*give money*
2	v_{act}	det:*a/an*	n_{sg}	*give a book*
3	v_{act}	det:*the*	n_{sg}	*give the book*
4	v_{act}	det:DEM	n_{sg}	*give this book*
5	v_{act}	det:POSS	n_{sg}	*give my book*
6	v_{act}	det:NULL	n_{pl}	*give books*
7	v_{act}	det:*the*	n_{pl}	*give the books*
8	v_{act}	det:DEM	n_{pl}	*give those books*
9	v_{act}	det:POSS	n_{pl}	*give my books*
10	v_{act}	det:OTHER	$n_{sg,pl}$	*give many books*
11	v_{pass}	det:ANY	$n_{sg,pl}$	*a/the/this/my book/books was/were given*

Afsaneh Fazly et al. 2009

Canonical Form

- Lexico-syntactic patterns that idiomatic usages tend to occur in

$$\mathcal{C}(v, n) = \{pt_k \in \mathcal{P} \mid z(v, n, pt_k) > T_z\}$$

$$z(v, n, pt_k) = \frac{f(v, n, pt_k) - \bar{f}}{s}$$

Afsaneh Fazly et al. 2009

Integrating Canonical Forms

- Unsupervised method used in Fazly et al. to identify canonical forms

- One-dimensional binary vector representing if the expression is in the canonical form

VNC-Tokens Dataset

Cook et al. 2008

- Dev
  - 14 MWEs
  - Training
    - 270 Idiom
    - 179 Literal
  - Testing
    - 92 Idiom
    - 53 Literal
- Test
  - 14 MWEs
  - Training
    - 298 Idiom
    - 172 Literal
  - Testing
    - 90 Idiom
    - 53 Literal

Accuracy

Model	DEV		TEST	
	−CF	+CF	−CF	+CF
CForm	-	0.721	-	0.749
Word2vec	**0.830**	**0.854**	**0.804**	**0.852**
Siamese CBOW	0.763	0.774	0.717	0.779
Skip-thoughts	0.803	0.827	0.786	0.842

Results per class

Model	Idiomatic			Literal		
	P	R	F	P	R	F
Word2vec −CF	0.815	0.879	0.830	0.627	0.542	0.556
Word2vec +CF	0.830	0.892	0.848	0.758	0.676	0.691

Conclusion

- Averaging word2vec embeddings outperforms all other models used
- Canonical form feature improves results
- Future work
  - Unseen MWEs
  - Other embedding models

Thank you

This work was financially supported by NSERC, NBIF, and University of New Brunswick

Results per class

Model	Idiomatic			Literal		
	P	R	F	P	R	F
CForm	0.766	0.901	0.794	0.668	0.587	0.576
Word2vec −CF	0.815	0.879	0.830	0.627	0.542	0.556
Word2vec +CF	0.830	0.892	0.848	0.758	0.676	0.691