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

Idiom Token Classification

- Verb-noun combination (VNC): Multiword expression consisting of a verb and a noun.
- Task: Determine if an instance of a verb-noun combination is idiomatic.

Idiomatic	Hereford United were <u>seeing stars</u> at Gillingham after letting in 2 early goals
Literal	Look into the night sky to <u>see</u> the <u>stars</u>

Models

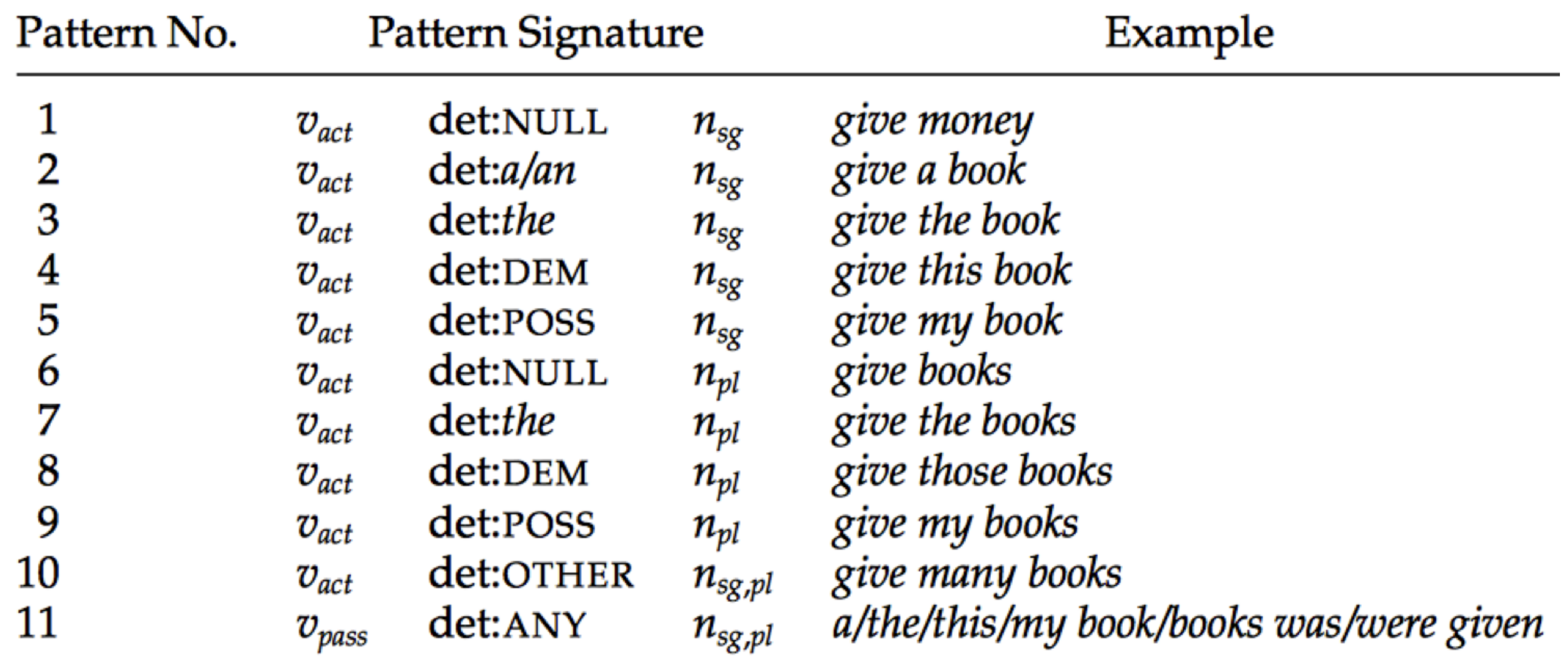
- Considered three approaches to representing VNC instances as a vector.
  - **1) Word2vec:** Average word2vec skipgram embeddings for the words in a sentence.
  - **2) Siamese CBOW:** Average siamese CBOW embeddings for the words in a sentence.
  - **3) Skip-thoughts:** Autoencoder that is trained to generate an embedding for a sentence using recurrent neural networks.
- **Supervised Model:** We trained an SVM for each of these representations.

Dataset

	DEV		TEST	
	Training	Testing	Training	Testing
MWE types	14	14	14	14
Idiom instances	270	92	298	90
Literal instances	179	53	172	53

Canonical Forms

- Lexico-syntactic patterns describing VNC instances.



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*

- Idiomatic VNC usages tend to occur in 1 canonical form.

Results

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

Conclusions and Future Work

- **Conclusions:**
  - Averaging word2vec embeddings outperforms the previously applied skip-thoughts model.
  - Employing information about canonical forms consistently improves all models.
- **Future Work:** Evaluate a word2vec model that is trained on the the same data as skip-thoughts.