Proceedings of the Society for Computation in Linguistics

Volume 2 Article 34

2019

What Do Neural Networks Actually Learn, When They Learn to Identify Idioms?

Marco Silvio Giuseppe Senaldi
Scuola Normale Superiore of Pisa, mrcsenaldi@gmail.com

Yuri Bizzoni
University of Gothenburg, yuri.bizzoni@gu.se

Alessandro Lenci
University of Pisa, alessandro.lenci@unipi.it

Recommended Citation

Senaldi, Marco Silvio Giuseppe; Bizzoni, Yuri; and Lenci, Alessandro (2019) "What Do Neural Networks Actually Learn, When They Learn to Identify Idioms?," *Proceedings of the Society for Computation in Linguistics*: Vol. 2 , Article 34.
DOI: https://doi.org/10.7275/x015-az15
Available at: https://scholarworks.umass.edu/scil/vol2/iss1/34

What do Neural Networks actually learn, when they learn to identify idioms?

Marco S. G. Senaldi[1], **Yuri Bizzoni**[2], **Alessandro Lenci**[3]
Scuola Normale Superiore - Italy[1], University of Gothenburg - Sweden[2], University of Pisa - Italy[3]
marco.senaldi@sns.it[1], yuri.bizzoni@gu.se[2], alessandro.lenci@unipi.it[3]

1 Introduction and Related Work

To date, Neural Networks (NNs) have been employed to carry out the automatic classification of various kinds of figurative expressions, like idioms (Bizzoni et al., 2017b) and metaphors (Do Dinh and Gurevych, 2016; Bizzoni et al., 2017a; Rei et al., 2017). It is common knowledge that metaphors (e.g., *my job is a jail*) reflect a transparent mapping from concrete examples in a source domain (e.g., the physical confinement of a jail) to abstract concepts in a target domain (e.g., the psychological constraints and tediousness of a job) (Lakoff and Johnson, 2008), while idioms (e.g., *buy the farm* 'to pass away', *shoot the breeze* 'to chat idly') synchronically appear as a rather heterogeneous class of semantically non-compositional multiword units that all in all exhibit greater lexicosyntactic rigidity, proverbiality and emotional valence with respect to literal expressions (Nunberg et al., 1994; Cacciari, 2014). In previous studies (Do Dinh and Gurevych, 2016; Rei et al., 2017), pre-trained word embeddings (Mikolov et al., 2013) have been fed to NNs to perform metaphor detection. Bizzoni et al. (2017a), for instance, successfully classify adjective-noun pairs where the same adjective is used either in a literal (e.g., *clean floor*) or a metaphorical (e.g., *clean performance*) sense with a neural classifier trained on a composition of the noun and adjective embeddings. As for idiom detection, Bizzoni et al. (2017b) use a fully-connected three-layered NN to automatically tell apart idiomatic and literal Italian verb-noun (VN) and adjective-noun (AN) phrases (e.g., *gettare la spugna* 'to throw in the towel' vs *vedere un film* 'to watch a movie; *alte sfere* 'high places' vs *nuova legge* 'new law'), training it with count-based vectors (Lenci, 2018) of the entire phrases taken as single tokens. Several works have nonetheless made it clear that it is still challenging to figure out the inner workings of NNs and the source of their performance (Karpathy et al., 2015), mostly because of their continuous representations and non-linearity that make it hard, for instance, to map their hidden states to interpretable language structures (Ding et al., 2017). By measuring the cosine similarity between the nouns in their dataset and the "metaphoricity vector" learnt by their network, Bizzoni et al. (2017a) found out that the algorithm was actually leveraging the concrete/abstract semantic shift undergone by the nouns while going from a literal to a metaphorical context. As the network performance in Bizzoni et al. (2017b) still remains unexplained, the aim of the present work was to shed light on which features in an idiom semantic vector are exploited by a NN when performing idiom vs literal classification, by means of an ablation paradigm (Greff et al., 2015; Kuncoro et al., 2016).

2 Our Proposal

Provided that the approach by Bizzoni et al. (2017b) uses just the count vectors of the idioms and literals to be distinguished as input, the aim of our work was to single out which semantic and contextual features are leveraged by the NN to carry out the classification task. Idioms are, as we stated above, a variegated class that, among the rest, displays varying levels of semantic ambiguity (Libben and Titone, 2008), i.e. whether a given idiom possesses a literal sense in addition to the figurative one (e.g., *spill the beans*, which can be both idiomatic and literal, vs *be on cloud nine*, which can be only idiomatic) and it is frequently used in that sense. On top of this, idioms, like metaphors, tend to be used to convey abstract concepts and are, generally speaking, less concrete in meaning with respect to literals (Citron et al., 2016). In the present research, we investigated whether seman-

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 310-313.
New York City, New York, January 3-6, 2019

tic ambiguity and concreteness might play a role in helping the NN to tell apart idioms and literals. In the ablation setting we implemented, we first tested our NN on the entire datasets of VN idiomatic and literal vectors, as in Bizzoni et al. (2017b). These will be called TOTAL models. In the so-called CONCRETENESS models we instead removed the most concrete literals from the training set so as to even out a difference in concreteness between the idioms and literals given as input to the NN. If the NN were actually relying on a difference in concreteness between idioms and literals to perform classification, we would expect the performance of these models to drop considerably. Finally, in the so-called AMBIGUITY models, we removed the most semantically ambiguous idioms from the training input. In our hypothesis, the fact that some idioms in the original dataset could have both a literal and an idiomatic meaning should be reflected in a richer and more variegated distributional representation with respect to expressions that can only receive a literal reading, and this could constitute a key factor to the neural classifier for spotting idioms. As in Bizzoni et al. (2017b), we employed pre-trained embeddings of our target idioms and literals taken as unanalyzed wholes, without composing the vectors of their component words. In light of idiom non-compositionality, Bizzoni et al. (2017b) have already shown models trained on vector composition to perform worse. Both Word2vec (Mikolov et al., 2013) and fastText (Bojanowski et al., 2017) vectors were used, in order to account for distributional information at both phrase and sub-phrase level. Finally, to assess whether our findings would hold crosslinguistically, we ran our models on two different datasets of Italian and English VN phrases respectively.

3 Datasets

3.1 Selection of the target expressions

Our Italian dataset was composed of 174 VN Italian idiomatic and literal expressions. First, a set of 87 Italian verbal idioms randomly chosen from idiom dictionaries (Quartu, 1993) was extracted from the itWaC corpus (Baroni et al., 2009;1,909M tokens ca). Their token frequency spanned from 63 (*parlare al muro* ‘to talk to a brick wall’) to 15,784 (*aprire le porte* ‘to open the floodgates’). Other 87 only-literal verbal phrases of comparable frequencies (e.g., *vedere un film* ‘to watch a movie’) were randomly selected. Our English dataset was instead composed of 120 VN idiomatic and literal expressions. From the COCA corpus (Davies, 2009; 520M tokens ca.) we extracted 60 English idioms, whose frequency spanned from 63 (*spill the beans*) to 1,641 tokens (*turn one’s back*), and other 60 only-literal phrases of comparable frequency (e.g., *eat a sandwich*).

3.2 Gold standard concreteness and ambiguity judgments

9 Italian linguistics students and researchers provided gold standard concreteness and ambiguity ratings for the Italian dataset. The 174 expressions were split into two sublists of 87 phrases. 3 raters per sublist evaluated how each phrase denoted an experience or concept related to one or more sensory modalities on a 1-7 Likert scale, with 1 standing for “totally abstract” and 7 standing for “totally concrete”. Other 3 judges were presented with the 87 idioms and voted on a 1-7 scale how plausible and frequent was to find each expression used in its literal sense in both written and spoken Italian, with 1 meaning “totally implausible” and 7 meaning “totally plausible”. Literals (M = 4.84) were rated as significantly more concrete than idioms (M = 3.16; W = 1887, $p < .001$), while 32 idioms out of 87 (36.78%) reported an average ambiguity score ≥ 5. 6 North American linguistics students and researchers rated the English dataset. The 120 idioms and literals were split into two sublists of 60 expressions, each of which was judged for concreteness by 2 subjects. Other 2 judges rated the 60 idioms for semantic ambiguity. Once again, literals (M = 6.20) were rated as significantly more concrete than idioms (M = 2.43; W = 76, $p < .001$). 31 idioms out of 60 (51.67%) got an average ambiguity score ≥ 5.

4 Method

4.1 Vector extraction

To represent our 174 Italian and 160 English idiomatic and literal VN constructions, we experimented with both Word2vec (Mikolov et al., 2013) and fastText (Bojanowski et al., 2017) vectors. We trained 300-dimensional embeddings with a Skip-gram model, using a symmetric window of 5 words and 10 negative examples. The vectors of the Italian expressions were trained on itWaC (Baroni et al., 2009), while the English ones were trained on COCA (Davies, 2009).

4.2 Training and test sets

In the TOTAL models, the entire sets of 174 Italian and 160 English items were randomly split into training and test sets roughly corresponding to the 80% and the 20% of the original sets respectively. 5 random splits were created for either dataset. In the CONCRETENESS models, we leveled the concreteness difference between idioms and literals in the training sets by removing all the literals with average concreteness > 5 and randomly trimming part of the idiom set to get an equal number of idioms and literals. A Kolmogorov-Smirnov test showed the distribution of concreteness judgments in the resulting Italian datasets to be not significantly different between idioms and literals. As for the English datasets, since concreteness ratings for literals were far higher than those given to idioms, a significant idioms-literals difference still remained, though we still removed the most concrete literals and the least concrete idioms. We finally assured that about 30% of the idioms still maintained an ambiguity score > 5, so as to disentangle the effects of concreteness and ambiguity. In the AMBIGUITY models, we removed from the input all the idioms with an average ambiguity ≥ 5, we randomly trimmed literals until we obtained and equal number of idioms and literals and we made sure via a Kolmogorov-Smirnov test that the distribution in concreteness judgments remained significantly different between idioms and literals. To sump up, 5 TOTAL, CONCRETENESS and AMBIGUITY datasets were randomly created for both Italian and English and fed to our NN.

4.3 The NN classifier

The NN we built was composed of three fully connected hidden layers.[1] The input layer has the same dimensionality of the original word embeddings and the output layer has dimensionality 1. The other two hidden layers have dimensionality 12 and 8. The network takes in input a single word embedding at a time. As said in Section 4.1, our embeddings had 300 dimensions each and encoded the distributional behavior of an entire phrase considered as a single token, without composing the vectors of its components. The most important dimensionality reduction is done by the first hidden layer, while the last layer applies a sigmoid activation function on the output to produce a binary judgment. In the classification task, we defined idioms as positive examples and non-idioms as negative examples of our training set.

5 Results

The average F1 scores of the ablated models and their SDs are reported in Table 1. Each F1 is averaged over 5 runs. The NN-based models of each dataset are compared with a RANDOM baseline. In both the Italian and the English dataset, while removing ambiguous idioms penalized the performance only marginally, though consistently, leveling the concreteness difference between idioms and literals led to much poorer results. Though the variation in performance across the runs (in terms of SD) was generally high for the CONCRETENESS models, the performance drop was nonetheless consistent with both datasets and vector types. Interestingly, CONCRETENESS models performed generally worse than the RANDOM baselines. The greater abstractness in meaning exhibited by idioms constitutes therefore a key element for our NN to perform idiom identification, while semantic ambiguity does not seem to be a determining factor. Finally, the kind of distributional information employed (Word2vec vs fastText) did not seem to impact the results.

6 Conclusions

In this ablation study we investigated which distributional and semantic features are leveraged by a NN to carry out idiom identification when it is just given phrase vectors as input. As it turns out, our NN was mostly exploiting a difference in concreteness rather than learning non-compositionality itself. From a more general standpoint, our findings suggest that when NNs are trained to spot a complex and multifaceted phenomenon such as idiomaticity, they rather exploit other underlying semantic features. Future work should investigate which other features to give in input to arrive at a more solid idiom-specific classification.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

[1] We used Keras, a library running on TensorFlow (Abadi et al., 2016).

Dataset	Model	Avg. training size	Test size	Avg. F1 (Word2vec)	SD	Avg. F1 (fastText)	SD
ITA	TOTAL	140 (70+70)	14 (37+37)	.78	.04	.71	.05
ITA	AMBIGUITY	87.2 (43.6+43.6)	14 (37+37)	.71	.06	.68	.06
ITA	CONCRETENESS	62 (31+31)	14 (37+37)	.41	.1	.40	.13
ITA	RANDOM		14 (37+37)	.44 (SD = .10)			
ENG	TOTAL	96 (48+48)	24 (12+12)	.65	.05	.64	.04
ENG	AMBIGUITY	47.6 (23.8+23.8)	24 (12+12)	.58	.15	.60	.09
ENG	CONCRETENESS	31.2 (15.6+15.6)	24 (12+12)	.33	0	.49	.21
ENG	RANDOM		24 (12+12)	.49 (SD = .10)			

Table 1: Average F1 scores and their SDs for the ablated models and the two random baselines. The sums between parentheses indicate the number of idioms+non-idioms in each set.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Yuri Bizzoni, Stergios Chatzikyriakidis, and Mehdi Ghanimifard. 2017a. "Deep" learning: Detecting metaphoricity in adjective-noun pairs. In *EMNLP 2017*.

Yuri Bizzoni, Marco S. G. Senaldi, and Alessandro Lenci. 2017b. Deep-learning the ropes: Modeling idiomaticity with neural networks. In *Fourth Italian Conference on Computational Linguistics CLiC-it 2017*, pages 36–41. Accademia University Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cristina Cacciari. 2014. Processing multiword idiomatic strings: Many words in one? *The Mental Lexicon*, 9(2):267–293.

Francesca MM Citron, Cristina Cacciari, Michael Kucharski, Luna Beck, Markus Conrad, and Arthur M Jacobs. 2016. When emotions are expressed figuratively: Psycholinguistic and affective norms of 619 idioms for German (PANIG). *Behavior research methods*, 48(1):91–111.

Mark Davies. 2009. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159. Association for Computational Linguistics.

Erik-Lân Do Dinh and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33.

Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. 2015. LSTM: A search space odyssey. *CoRR*, abs/1503.04069.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Adhiguna Kuncoro, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig, and Noah A Smith. 2016. What do recurrent neural network grammars learn about syntax? *arXiv preprint arXiv:1611.05774*.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Alessandro Lenci. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171.

Maya R Libben and Debra A Titone. 2008. The multi-determined nature of idiom processing. *Memory & Cognition*, 36(6):1103–1121.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, volume 13, pages 746–751.

Geoffrey Nunberg, Ivan Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.

Monica B. Quartu. 1993. *Dizionario dei modi di dire della lingua italiana*. RCS Libri.

Marek Rei, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. *arXiv preprint arXiv:1709.00575*.