Deep-learning the Ropes: Modeling Idiomaticity with Neural Networks

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

Abstract

English. In this work we explore the possibility of training a neural network to classify and rank idiomatic expressions under constraints of data scarcity. We discuss our results comparing them both to other unsupervised models designed to perform idiom detection and to similar supervised classifiers trained to detect metaphoric bigrams.

Italiano. *In questo lavoro esploriamo la possibilità di addestrare una rete neurale per classificare ed ordinare espressioni idiomatiche in condizioni di scarsità di dati. I nostri risultati sono discussi in comparazione sia con altri algoritmi non supervisionati ideati per l'identificazione di espressioni idiomatiche sia con classificatori supervisionati dello stesso tipo addestrati per identificare bigrammi metaforici.*

1 Introduction

Figurative expressions like idioms (e.g. *to learn the ropes* 'to learn how to do a job', *to cut the mustard* 'to perform up to expectations', etc.) and metaphors (e.g. *clean performance*, *that lawyer is a shark*, etc.) are pervasive in language use. Important differences have been stressed between the two types of expressions from a theoretical (Gibbs, 1993; Torre, 2014), neurocognitive (Bohrn et al., 2012) and corpus linguistic (Liu, 2003) prespective. On the one hand, as stated by Lakoff and Johnson (2008), linguistic metaphors reflect an instantiation of conceptual metaphors, whereby abstract concepts in a *target domain* (e.g. the ruthlessness of a lawyer) are described by a rather transparent mapping to concrete examples taken from a *source domain* (e.g. the aggressiveness of a shark). On the other hand, although most idioms originate as metaphors (Cruse, 1986), they have undergone a crystallization process in diachrony, whereby they now appear as fixed and non-compositional word combinations that belong to the wider class of Multiword Expressions (MWEs) (Sag et al., 2002) and always exhibit lexical and morphosyntactic rigidity to some extent (Cacciari and Glucksberg, 1991; Nunberg et al., 1994). It is anyway crucial to underline that idiomaticity itself is a multidimensional and gradient phenomenon (Nunberg et al., 1994; Wulff, 2010) with different idioms showing varying degrees of semantic transparency, formal versatility, proverbiality and affective valence.

The aim of this work is to explore the fuzzy boundary between idiomatic and metaphorical expression, by applying a method designed to discriminate figurative vs. literal usages to the task of distinguishing idiomatic from compositional expressions. Our starting point is the work of Bizzoni et al. (2017). The authors managed to classify adjective-noun pairs where the same adjectives were used both in a metaphorical and a literal sense (e.g. *clean performance* vs. *clean floor*) using a neural classifier trained on a composition of the words' embeddings (Mikolov et al., 2013). Actually, the neural network was able to detect the abstract/concrete semantic shift of nouns when used with the same adjective in figurative and literal compositions respectively, basically treating the noun as the "context" to discriminate the metaphoricity of the adjective. In our attempt, we will use a relatively similar approach to classify idiomatic expressions by training a three-layered neural network on a set of idiomatic and non-idiomatic expressions and we'll compare the performance of the network when trained on different syntactic patterns (Adjective-Noun and Verb-Noun expressions, AN and VN henceforth).

Importantly, the abstract/concrete polarity the

network was able to learn in Bizzoni et al. (2017) will not be available this time, since none of the idiom constituents will ever appear in its literal sense inside the expressions, whatever their concreteness may be. What we want to find out is whether the sole information captured by the distributional vector of a single expression is sufficient to learn its potential idiomaticity. Differently from Bizzoni et al. (2017), for each idiom we collect a count-based vector (Turney and Pantel, 2010) of the expression as a whole, taken as a single token. We compare this approach with a model trained on the composition of the individual words of an expression, showing that the latter is less effective for idioms than for metaphors. In both cases we will be operating on scarce training sets (26 AN and 90 VN constructions). Traditional ways to deal with data scarcity in computational linguistics resort to a wide number of different features to annotate the training set (see for example Tanguy et al. (2012)) or rely on artificial bootstrapping of the training set (He and Liu, 2017). In our case we test the performance of our classifier on scarce data without bootstrapping the dataset and relying only on the information provided by the distributional semantic space, showing that the distribution of an expression in large corpora can provide enough information to learn idiomaticity from few examples with a satisfactory degree of accuracy.

2 Related Work

Previous computational research has exploited different methods to perform *idiom type detection* (i.e., automatically telling apart potential idioms like *to get the sack* from only literal combinations like *to kill a man*). For example Lin (1999) and Fazly et al. (2009) label a given word combination as idiomatic if the Pointwise Mutual Information (PMI) (Church and Hanks, 1991) between its constituents is higher than the PMIs between the components of a set of lexical variants of this combination obtained by replacing the component words of the original expressions with semantically related words. Other studies have resorted to Distributional Semantics (Lenci, 2008; Turney and Pantel, 2010) by measuring the cosine between the vector of a given phrase and the single vectors of its components (Fazly and Stevenson, 2008) or between the phrase vector and the sum or product vector of its components (Mitchell and Lapata, 2010; Krčmář et al., 2013). Senaldi et al. (2016b) and Senaldi et al. (2016a) have combined insights from both these approaches by observing that the vectors of VN and AN idioms are less similar to the vectors of lexical variants of these expressions with respect to the vectors of compositional constructions. To the best of our knowledge, neural networks have been previously adopted to perform MWE detection in general (Legrand and Collobert, 2016; Klyueva et al., 2017), but not idiom identification specifically. In Bizzoni et al. (2017), pre-trained noun and adjective vector embeddings are fed to a single-layered neural network to disambiguate metaphorical and literal AN combinations. Several combination algorithms are experimented with to concatenate adjective and noun embeddings. All in all, the method is shown to outperform the state of the art, presumably leveraging the abstractness degree of the noun as a clue to metaphoricity.

3 Dataset

3.1 Target expressions extraction

The two idiom datasets we employ in the current study come from Senaldi et al. (2016b) and Senaldi et al. (2016a). The first one is composed of 45 idiomatic and 45 non-idiomatic Italian V-NP and V-PP constructions (e.g. *tagliare la corda* 'to flee' lit. 'to cut the rope' and *leggere un libro* 'to read a book') that were selected from an Italian idiom dictionary (Quartu, 1993) and extracted from the itWaC corpus (Baroni et al., 2009), composed of about 1,909M tokens. Their frequency spanned from 364 (*ingannare il tempo* 'to while away the time') to 8294 (*andare in giro* 'to get about'). The latter comprises 13 idiomatic and 13 non-idiomatic AN constructions (e.g. *punto debole* 'weak point' and *nuova legge* 'new law') that were still extracted from itWaC and whose frequency varied from 21 (*alte sfere* 'high places', lit. 'high spheres') to 194 (*punto debole*).

3.2 Building target vectors

Count-based Distributional Semantic Models (DSMs) (Turney and Pantel, 2010) allow for representing words and expressions as high-dimensionality vectors, where the vector dimensions register the co-occurrence of the target words or expressions with some contextual features, e.g. the content words that linearly precede and follow the target element within a fixed contextual window. We built two DSMs on itWaC, where our tar-

get AN and VN idioms and non-idioms were represented as target vectors and co-occurrence statistics counted how many times each target construction occurred in the same sentence with each of the 30,000 top content words in the corpus. Differently from Bizzoni et al. (2017), we did not opt for prediction-based vector representations (Mikolov et al., 2013). Although some studies have brought out that context-predicting models fare better than count-based ones on a variety of semantic tasks (Baroni et al., 2014), including compositionality modeling (Rimell et al., 2016), others (Blacoe and Lapata, 2012; Cordeiro et al., 2016) have shown them to perform comparably. Moreover, Levy et al. (2015) highlight that much of the superiority in performance exhibited by word embeddings is actually due to hyperparameter optimizations, which, if applied to traditional models as well, can bring to equivalent outcomes. Therefore, we felt confident in resorting to count-based vectors as an equally reliable representation for the task at hand.

3.3 Gold standard idiomaticity judgments

In Senaldi et al. (2016b) and Senaldi et al. (2016a), we collected gold standard idiomaticity judgments for our target AN and VN constructions. 9 Linguistics students were presented with a list of our 26 AN constructions and were asked to evaluate how idiomatic each expression was from 1 to 7, with 1 standing for 'totally compositional' and 7 standing for 'totally idiomatic'. Inter-coder agreement, measured with Krippendorff's α (Krippendorff, 2012), was equal to 0.76. The same procedure was repeated for our 90 VN constructions, but in this case the inital list was split into 3 sublists of 30 expressions, each one to be rated by 3 subjects. Krippendorff's α was 0.83 for the first sublist and 0.75 for the other two.

4 Classifier

We built a neural network composed of three "dense" or fully connected layers[1] of dimensionality 12, 8 and 1 respectively. Our network takes in input a single vector at a time, which can be a word embedding, a count-based distributional vector or a composition of several word vectors. For the core part of our experiment we used as input single distributional vectors of two-word expressions. Due to our input's magnitude, the most important reduction of data dimensionality is carried out by the first layer of our model. The last layer applies a sigmoid activation function on the output in order to produce a binary judgment. While binary scores are necessary to compute the model classification accuracy and will be evaluated in terms of F1, our model's continuous scores can be retrieved and will be used to perform an ordering task on the test set, that we will evaluate in terms of Interpolated Average Precision (IAP) [2] and against the human idiomaticity judgments with Spearman's ρ.

5 Evaluation

We trained our model on the 30,000 dimensional distributional vectors of VN and AN expressions as well as on the composition of their individual words' vectors. We tried with different semantic spaces as well. When trained on PPMI- (Church and Hanks, 1991) and SVD-transformed (Deerwester et al., 1990) vectors of 150, 200, 250 and 300 dimensions, our models performed comparably or even worse; so, results for these cases won't be presented here. Details of both classification and ordering task are shown in Table 1.

5.1 Verb-Noun

We ran our model on the VN dataset, composed of 90 elements, 45 idioms and 45 non-idiomatic expressions. This is the larger of the two datasets. We trained our model both on 30 and 40 elements for 20 epochs and tested on the remaining 60 and 50 elements respectively, reaching a maximum IAP of 0.87 and Spearman's ρ of 0.76. In general we found the model's performance, both in accuracy and in correlation, comparable to the results reported in Senaldi et al. (2016b), who reached a maximum IAP of 0.91 and a maximum Spearman's ρ of -0.67.

5.2 Adjective-Noun

We ran our model on the AN dataset, composed of 26 elements, 13 idioms and 13 non-idiomatic expressions. We empirically found that our model was able to perform some generalization on the data when the training set contained at least 14 elements, evenly balanced between positive and negative examples. We trained our model on 16 elements for 30 epochs and tested on the remaining 10 elements. While accuracy's exact value can

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

[2] Following Fazly et al. (2009), IAP was computed at recall levels of 20%, 50% and 80%.

Vector	Training	Test	IAP	rho	F1
VN	15+15	30+30	0.82	0.50***	0.8
VN	20+20	15+15	0.82	0.76***	0.87
Concat (VN)	15+15	14+14	0.7	0.47*	0.69
AN	8+8	6+4	1?	0.93***	0.9
VN+AN	23+23	14+14(VN)	0.9	0.76***	0.82
VN+AN	23+23	18+20(joint)	0.8	0.64***	0.76
VN+AN	23+23	5+5(AN)	0.57	-0.31	0.58

Table 1: Interpolated Average Precision, Spearman's correlation with the speaker judgments and F-measure for Vector-Noun training (VN), Adjective-Noun training (AN), joint training and training through vector concatenation (** = $p < .01$, *** = $p < .001$). Training and test set are expressed as the sum of positive and negative examples.

undergo some fluctuations when a model is trained on very small sets, we always registered accuracies higher than 80%, with 4 out of 5 idioms correctly labeled in every trial. We reached an IAP of 1.0 and a ρ of 0.93, although it is important to keep in mind that such scores are computed on a very restricted test set. Senaldi et al. (2016b) reached a maximum IAP of 0.85 and a maximum ρ of -0.68. When the training size was under the critical threshold, accuracy dropped significantly. With training sets of 10 or 12 elements, our model naturally went in overfitting, quickly reaching 100% accuracy on the training set and failing to correctly classify unforeseen expressions. In these cases a partial learning was still visible in the ordering task, where most idioms, even if labeled incorrectly, received higher scores than non-idioms.

5.3 Joint training

We also tried to train our model on both datasets together, to check to what extent it would be able to recognize the same underlying semantic phenomenon through different syntactic constructions. We used two different approaches for this experiment. Training our model first on one dataset, e.g. the AN pairs, and then on the other required more epochs overall (more than 100) to stabilize and resulted in a poorer performance (66% F-measure on both test sets). Training our model on a mixed dataset containing the elements of both training sets, our model employed only 12 epochs to reach an F-measure of 76% on the mixed training set. Anyway, we also noticed that VN expressions were learned better than AN expressions. In short, our model was able to generalize over the two datasets, but this involved a loss in accuracy.

5.4 Vector composition

In addition to using the vector of an expression as a whole, we tried to feed our model with the concatenation of the vectors of the single words in an expression, as in Bizzoni et al. (2017). For example, instead of using the 30,000 dimensional vector of the expression *cambiare musica*, we used the 60,000 dimensional vector resulting from the concatenation of *cambiare* and *musica*. We ran this experiment only on the VN dataset, being the largest and the one that yielded the best results in the previous settings. We used 30 elements in training and 26 in testing and trained our model for 80 epochs overall. Predictably enough, vector composition resulted in the worst performance, differently from what happened with metaphors (Bizzoni et al., 2017); nonetheless, the results are not completely random: with an F1 of 69%, the model seems able to learn idiomaticity to a lower, but not null, degree; these findings would be in line with the claim that the meaning of the subparts of several idioms, while less important than in metaphors, is not completely obliterated (McGlone et al., 1994).

6 Error Analysis

Two frequent false positives are *tagliare il traguardo* and *abbassare la guardia*. While we labeled them as non-idioms in our dataset, since they're rather compositional, nonetheless they can be very often used figuratively and that's probably why our algorithms identified them as idioms. A frequent false negative was *vedere la luce*, which probably occurs more often in its literal sense in the corpus we used.

7 Discussion and Conclusions

It seems that the distribution of idiomatic and compositional expressions in large corpora can suffice for a supervised classifier to learn the difference between the two linguistic elements from small training sets and with a good level of accuracy. Unlike with metaphors (Bizzoni et al., 2017), feeding the classifier with a composition of the individual words' vectors of such expressions performs quite scarcely and can be used to detect only some idioms. This takes us back to the core difference that while metaphors are more compositional and preserve a transparent source domain to target domain mapping, idioms are by and large non-compositional. Since our classifiers rely only on contextual features, their ability in classification must stem from a difference in distribution between idioms and non-idioms. A possible explanation is that while the literal expressions we selected, like *vedere un film* or *ascoltare un discorso*, tend to be used with animated subjects and thus to appear in more concrete contexts, most of our idioms (e.g. *cadere dal cielo* or *lasciare il segno*) allow for varying degrees of animacy or concreteness of the subject, and thus their context can easily get more diverse. At the same time, the drop in performance we observe in the joint models seems to indicate that the different parts of speech composing our elements entail a significant contextual difference between the two groups, which introduces a considerable amount of uncertainty in our model. It is also possible that other contextual elements we did not consider have played a role in the learning process of our models. We intend to deepen this aspect in future works.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

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

Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 238–247.

Bizzoni, Y., Chatzikyriakidis, S., and Ghanimifard, M. (2017). "Deep" learning: Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52.

Blacoe, W. and Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 546–556. Association for Computational Linguistics.

Bohrn, I. C., Altmann, U., and Jacobs, A. M. (2012). Looking at the brains behind figurative language: A quantitative meta-analysis of neuroimaging studies on metaphor, idiom, and irony processing. *Neuropsychologia*, 50(11):2669–2683.

Cacciari, C. and Glucksberg, S. (1991). Understanding idiomatic expressions: The contribution of word meanings. *Advances in Psychology*, 77:217–240.

Church, K. W. and Hanks, P. (1991). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Cordeiro, S., Ramisch, C., Idiart, M., and Villavicencio, A. (2016). Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1986–1997.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391.

Fazly, A., Cook, P., and Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 1(35):61–103.

Fazly, A. and Stevenson, S. (2008). A distributional account of the semantics of multiword

expressions. *Italian Journal of Linguistics*, 1(20):157–179.

Gibbs, R. W. (1993). Why idioms are not dead metaphors. *Idioms: Processing, structure, and interpretation*, pages 57–77.

He, X. and Liu, Y. (2017). Not enough data?: Joint inferring multiple diffusion networks via network generation priors. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 465–474.

Klyueva, N., Doucet, A., and Straka, M. (2017). Neural networks for multi-word expression detection. *Proceedings of the 13th Workshop on Multiword Expressions*, pages 60–65.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology*. Sage.

Krčmář, L., Ježek, K., and Pecina, P. (2013). Determining Compositionality of Expresssions Using Various Word Space Models and Measures. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 64–73.

Lakoff, G. and Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Legrand, J. and Collobert, R. (2016). Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 67–71.

Lenci, A. (2008). Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.

Levy, O., Goldberg, Y., and Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Lin, D. (1999). Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324.

Liu, D. (2003). The most frequently used spoken american english idioms: A corpus analysis and its implications. *Tesol Quarterly*, 37(4):671–700.

McGlone, M. S., Glucksberg, S., and Cacciari, C. (1994). Semantic productivity and idiom comprehension. *Discourse Processes*, 17(2):167–190.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26tth International Conference on Neural Information Processing System*, pages 3111–3119.

Mitchell, J. and Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.

Nunberg, G., Sag, I., and Wasow, T. (1994). Idioms. *Language*, 70(3):491–538.

Quartu, M. B. (1993). *Dizionario dei modi di dire della lingua italiana*. RCS Libri.

Rimell, L., Maillard, J., Polajnar, T., and Clark, S. (2016). RELPRON: A relative clause evaluation data set for compositional distributional semantics. *Computational Linguistics*, 42(4):661–701.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15.

Senaldi, M. S. G., Lebani, G. E., and Lenci, A. (2016a). Determining the compositionality of noun-adjective pairs with lexical variants and distributional semantics. In *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it 2016)*, pages 268–273.

Senaldi, M. S. G., Lebani, G. E., and Lenci, A. (2016b). Lexical variability and compositionality: Investigating idiomaticity with distributional semantic models. In *Proceedings of the 12th Workshop on Multiword Expression*, pages 21–31.

Tanguy, L., Sajous, F., Calderone, B., and Hathout, N. (2012). Authorship attribution: Using rich linguistic features when training data is scarce. In *PAN Lab at CLEF*.

Torre, E. (2014). *The emergent patterns of Italian idioms: A dynamic-systems approach*. PhD thesis, Lancaster University.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Wulff, S. (2010). *Rethinking Idiomaticity: A Usage-based Approach*. A&C Black.