Finding the Neural Net: Deep-learning Idiom Type Identification from Distributional Vectors

Yuri Bizzoni*
University of Gothenburg

Marco S. G. Senaldi**
Scuola Normale Superiore di Pisa

Alessandro Lenci†
Università di Pisa

The present work aims at automatically classifying Italian idiomatic and non-idiomatic phrases with a neural network model under constrains of data scarcity. Results are discussed in comparison with an existing unsupervised model devised for idiom type detection and a similar supervised classifier previously trained to detect metaphorical bigrams. The experiments suggest that the distributional context of a given phrase is sufficient to carry out idiom type identification to a satisfactory degree, with an increase in performance when input phrases are filtered according to human-elicited idiomaticity ratings collected for the same expressions. Crucially, employing concatenations of single word vectors rather than whole-phrase vectors as training input results in the worst performance for our models, differently from what was previously registered in metaphor detection tasks.

1. Introduction

Generally speaking, figurativeness has to do with pointing at a contextual interpretation for a given expression that goes beyond its mere literal meaning (Frege 1892; Gibbs et al. 1997; Cacciari and Papagno 2012). Let's imagine a commentator that, referring to an athlete, says *She's always delivered clean performances but this one really took the cake!* In this sentence, *clean performances* is an example of *metaphorical expression* that, according to the model proposed by Lakoff and Johnson (2008), reflects a rather transparent mapping between an abstract concept in a *target domain* (e.g., the flawlessness of a performance) and a concrete example taken from a *source domain* (e.g., the cleanliness of a surface). On the other hand, *take the cake* is an *idiom*, i.e. a lexicosyntactically rigid multiword unit (Sag et al. 2002) that is entirely non-compositional, since its meaning of 'being outstanding' is not accessible by simply composing the meanings of *take* and *cake* and must therefore be learnt by heart by speakers (Frege 1892; Cacciari 2014).

Important differences have been stressed between metaphors and idioms in theoretical (Gibbs 1993; Torre 2014), neurocognitive (Bohrn, Altmann, and Jacobs 2012) and corpus linguistic (Liu 2003) studies. First of all, metaphors represent a productive

* Department of Philosophy, Linguistics, Theory of Science - Dicksonsgatan 4, 41256, Göteborg, Sweden. E-mail: yuri.bizzoni@gu.se
** Scuola Normale Superiore - Piazza dei Cavalieri 7, I-56126 Pisa, Italy. E-mail: marco.senaldi@sns.it
† CoLing Lab, Department of Philology, Literature and Linguistics - Via S. Maria 36, I-56126 Pisa, Italy. E-mail: alessandro.lenci@unipi.it

phenomenon: studies on metaphor production strategies indeed show a large ability of language users to generalize and create new metaphors on the fly from existing ones, allowing researchers to hypothesize recurrent semantic mechanisms underlying a large number of productive metaphors (McGlone 1996; Lakoff and Johnson 2008). For example, starting from the *clean performance* metaphor above, we could also say the delivered performance was *neat*, *spick-and-span* and *crystal-clear* by sticking to the same conceptual domain of cleanliness. 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 conventionalized and (mostly) fixed combinations that form a finite repository in a given language (Nunberg, Sag, and Wasow 1994). From a formal standpoint, though some idioms allow for restricted lexical variability (e.g., the concept of getting crazy can be conveyed both by *to go nuts* and *to go bananas*), this kind of variation is not as free and systematic as with metaphors and literal language (e.g., transforming the *take the cake* idiom above into *take the candy* would hinder a possible idiomatic reading) (Fraser 1970; Geeraert, Baayen, and Newman 2017). From the semantic point of view, it is interesting to observe how speakers can correctly use the most semantically opaque idioms in discourse without necessarily being aware of their actual metaphorical origin or anyway having contrasting intuitions about it. For example, Gibbs (1994) reports that many English speakers explain the idiom *kick the bucket* 'to die' as someone kicking a bucket to hang themselves, while it actually originates from a corruption of the French word *buquet* indicating the wooden framework that slaughtered hogs kicked in their death struggles. Secondly, metaphorical expressions can receive varying interpretations according to the context at hand: saying that *John is a shark* could mean that he's ruthless on his job, that he's aggressive or that he attacks people suddenly (Cacciari 2014). Contrariwise, idiomatic expressions always keep the same meaning: saying that *John kicked the bucket* can only be used to state that he passed away. Finally, idioms and metaphors differ in the mechanisms they recruit in language processing: while metaphors seem to bring into play *categorization* (Glucksberg, McGlone, and Manfredi 1997) or *analogical* (Gentner 1983) processes between the vehicle and the topic (e.g., *shark* and *John* respectively in the sentence above), idioms by and large call for lexical access mechanisms (Cacciari 2014). Nevertheless, it is crucial to underline that idiomaticity itself is a multidimensional and gradient phenomenon (Nunberg, Sag, and Wasow 1994; Wulff 2008) with different idioms showing varying degrees of semantic transparency, formal versatility, proverbiality and affective valence. All this variance within the class of idioms themselves has been demonstrated to affect the processing of such expressions in different ways (Cacciari 2014; Titone and Libben 2014).

The aim of this work is to focus on the fuzzy boundary between idiomatic and metaphorical expressions from a computational viewpoint, by applying a supervised method previously designed to discriminate metaphorical vs. literal usages of input constructions to the task of distinguishing idiomatic from compositional expressions. Our starting point is the work of Bizzoni, Chatzikyriakidis, and Ghanimifard (2017), who 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*) by means of a neural classifier trained on a composition of the words' embeddings (Mikolov et al. 2013). As the authors found out, the neural network succeeded in the task because it was able to detect the abstract/concrete semantic shift undergone by the nouns when used with the same adjective in figurative and literal compositions respectively. 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 Italian idioms (e.g. *gettare la spugna* 'to

throw in the towel', lit. 'to throw the sponge') and non-idioms (e.g. *vedere una partita* 'to watch a match'). The performance of the network will be compared when trained with constructions belonging to different syntactic patterns, namely Adjective-Noun and Verb-Noun expressions (AN and VN henceforth). Noteworthily, the abstract/concrete polarity the network was able to learn in Bizzoni, Chatzikyriakidis, and Ghanimifard (2017) will not be available this time: while the nouns in the dataset of Bizzoni, Chatzikyriakidis, and Ghanimifard (2017) were used in their literal sense, idioms are entirely non-compositional, so none of their constituents is employed literally inside the expressions, independently of their concreteness (e.g., *spugna* 'sponge' in *gettare la spugna* vs *numeri* 'numbers' in *dare i numeri* 'to lose it', lit. 'to give the numbers'). What we want to find out is whether the sole information captured by the distributional vector of a given expression is sufficient for the network to learn its potential idiomaticity. The idiom classification scores of our models will be compared with those obtained by Senaldi, Lebani, and Lenci (2016) and Senaldi, Lebani, and Lenci (2017), who propose a distributional semantic algorithm for idiom type detection. Our study employs their small datasets. Therefore, the training sets we will operate on will be very scarce. 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.

This paper is structured as follows: after reviewing in Section 2 the existing literature on idiom and metaphor processing, in Section 3 we will briefly outline the experimental design and in Section 4 we will provide details about the dataset we used and the human ratings we collected to validate our algorithms; in Section 5 we will go through the structure and functioning of our classifier and in Section 7 we will evaluate the performance of our models. Section 8 presents a qualitative error analysis, then followed by a discussion of the results (Section 9).

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, Cook, and Stevenson (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, 2018; 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, Ježek, and Pecina 2013). Senaldi, Lebani, and Lenci (2016) and Senaldi, Lebani, and Lenci (2017) combine insights from both these approaches. They start from two lists of 90 VN and 26 AN constructions, the former composed of 45 idioms (e.g., *gettare la spugna*) and 45 non-idioms (e.g., *vedere una partita*), the latter comprising 13 idioms (e.g., *filo rosso* 'common thread', lit. 'red thread') and 13 non-idioms (e.g., *lungo periodo*

'long period'). For each of these constructions, a series of lexical variants are generated distributionally or via MultiWordNet (Pianta, Bentivogli, and Girardi 2002) by replacing the subparts of the constructions with semantically related words (e.g. from *filo rosso*, variants like *filo nero* 'black thread', *cavo rosso* 'red cable' and *cavo nero* 'black cable' are generated). What comes to the fore is that the vectors of the idiomatic expressions are less similar to the vectors of their lexical variants with respect to the similarity between the vector of a literal constructions and the vectors of its lexical alternatives. To provide an example, the cosine similarity between the vector of an idiom like *filo rosso* and the vectors of its lexical variants like *filo nero* and *cavo rosso* was found to be smaller than the cosine similarity between the vector of a literal phrase like *lungo periodo* and the vectors of its variants like *interminabile tempo* 'endless time' and *breve periodo* 'short period'.

Moving to the methodology exploited in the current study, to the best of our knowledge, neural networks have been previously adopted to perform MWE detection in general (Legrand and Collobert 2016; Klyueva, Doucet, and Straka 2017), but not idiom identification specifically. As mentioned in the Introduction, in Bizzoni, Chatzikyriakidis, and Ghanimifard (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 figurativeness and basically treating the noun as the "context" to discriminate the metaphoricity of the adjective (cf. *clean performance* vs *clean floor*, where *performance* is more abstract than *floor* and therefore the mentioned cleanliness is to be intended metaphorically).

Besides Bizzoni, Chatzikyriakidis, and Ghanimifard (2017), using neural networks for metaphor detection with pretrained word embeddings initialization has been tried in a small number of recent works, proving that this is a valuable strategy to predict metaphoricity in datasets. Rei et al. (2017) present an ad-hoc neural design able to compose and detect metaphoric bigrams in two different datasets. Do Dinh and Gurevych (2016) apply a series of perceptrons to the VU Amsterdam Metaphor Corpus (Steen et al. 2014) combined with word embeddings and part-of-speech tagging. Finally, a similar approach - a combination of fully connected networks and pre-trained word embeddings - has also been used as a pre-processing step to metaphor detection, in order to learn word and sense abstractness scores to be used as features in a metaphor identification pipeline (Köper and Schulte im Walde 2017).

3. Method

In this work we carried out a supervised idiom type identification task by resorting to a three-layered neural network classifier. After selecting our dataset of VN and AN target expressions (Section 4.1), for which gold standard idiomaticity ratings had already been collected (Section 4.2), we built count vector representations for them (Section 4.3) from the itWaC corpus (Baroni et al. 2009) and fed them to our classifier (Section 5) with different training splits (Section 6). The network returned a binary output, whereby idioms were taken as our positive examples and non-idioms as our negative ones. Differently from Bizzoni, Chatzikyriakidis, and Ghanimifard (2017), for each idiom or non-idiom we initially built a count-based vector (Turney and Pantel 2010) of the expression as a whole, taken as a single token. We then compared this approach with a model trained on the concatenation of the individual words of an expression, but the latter turned out to be less effective for idioms than for metaphors. Each model was finally evaluated

in terms of classification accuracy, ranking performance and correlation between its continuous scores and the human-elicited idiomaticity judgments (Section 7).

Since we mostly worked with vectors that took our target expressions as unanalyzed wholes, as if they were single tokens, we were not concerned with the fact that some verbs were shared by more than one idiom (e.g., *lasciare il campo* 'to leave the field' and *lasciare il segno* 'to leave one's mark') or non-idiom (e.g., *andare a casa* 'to go home' and *andare all'estero* 'to go abroad') at once, given that our network could not access this information.

4. Dataset

4.1 Target expressions selection

The two datasets we employed in the current study come from Senaldi, Lebani, and Lenci (2016) and Senaldi, Lebani, and Lenci (2017). The first one is composed of 45 idiomatic Italian V-NP and V-PP constructions (e.g., *tagliare la corda* 'to flee' lit. 'to cut the rope') that were selected from an Italian idiom dictionary (Quartu 1993) and extracted from the itWaC corpus (Baroni et al. (2009), 1,909M tokens ca.) and whose frequency spanned from 364 (*ingannare il tempo* 'to while away the time') to 8294 (*andare in giro* 'to get about'), plus other 45 Italian non-idiomatic V-NP and V-PP constructions of comparable frequencies (e.g., *leggere un libro* 'to read a book'). The latter dataset 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*).

4.2 Gold standard idiomaticity judgments

Senaldi, Lebani, and Lenci (2016) and Senaldi, Lebani, and Lenci (2017) collected gold standard idiomaticity judgments for the 26 AN and 90 VN target constructions in their datasets. Nine linguistics students were presented with a list of the 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 the 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. These inter-coder agreement scores were taken as a confirmation of reliability for the collected ratings (Artstein and Poesio 2008). As will become clear in Section 6, these judgments served the twofold purpose of evaluating the classification performance of our neural network and filtering the expressions to use as training input for our models.

4.3 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 trained two DSMs on itWaC, where our target 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, Chatzikyriakidis, and Ghanimifard (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, Dinu, and Kruszewski 2014), including compositionality modeling (Rimell et al. 2016), others (Blacoe and Lapata 2012; Cordeiro et al. 2016) have shown them to perform comparably. In phrase similarity and paraphrase tasks, Blacoe and Lapata (2012) find count vectors to score better then or comparably to predict vectors built following Collobert and Weston (2008)'s neural language model. Cordeiro et al. (2016) show PPMI-weighted count-based models to perform comparably to *word2vec* (Mikolov, Yih, and Zweig 2013) in predicting nominal compound compositionality. Moreover, Levy, Goldberg, and Dagan (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.

5. The neural network classifier

We built a neural network composed of three "dense" or fully connected hidden layers.[1] The input layer has the same dimensionality of the original vectors and the output layer has dimensionality 1. The other two hidden layers have dimensionality 12 and 8 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. As we discussed in the previous section, these vectors have 30,000 dimensions each and represent the distributional behavior of a full expression rather than that of the individual words composing such expression. Given this distributional matrix, we defined idioms as positive examples and non-idioms as negative examples of our training set. Due to the magnitude of our input, the most important reduction of data dimensionality is carried out by the first hidden 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 Spearman's ρ with the human-elicited idiomaticity judgments. IAP and ρ, therefore, will be useful to investigate how good our model is in ranking idioms before non-idioms.

6. Choosing the training set

The scarcity of our training sets constitutes a challenge for neural models, typically designed to deal with massive amounts of data. The typical effect of such scarcity is a fluctuation in performance: training our model on two different sections of the same dataset is likely to result in quite different F-scores.

1 We used Keras, a library running on TensorFlow (Abadi et al. 2016).
2 Following Fazly, Cook, and Stevenson (2009), IAP was computed at recall levels of 20%, 50% and 80%.

Unless otherwise specified, the IAP, Spearman's ρ and F1 scores reported in Table 1 are averaged on 5 runs of each model on the same datasets: at each run, the training split is randomly selected. We found that some samples of the training set seemingly make it harder for the model to learn idiom detection. When such runs are included in the mean, the performance is drastically lowered.

In our attempt to understand whether we could find a rationale behind this phenomenon or it was instead completely unpredictable, in some versions of our models we have tried to filter our training sets according to the idiomaticity judgments we elicited from speakers (Section 4.2) to assess which composition of our training sets made our algorithm more effective. In the first approach, which we will label as High-to-Low (HtL henceforth), the network was trained on the idioms receiving the highest idiomaticity ratings (and symmetrically on the compositional expressions having the lowest idiomaticity ratings) and was therefore tested on the intermediate cases. In the second approach, which we called Low-to-High (LtH), the model was trained on more borderline exemplars, i.e. the idioms having the lowest idiomaticity ratings and the compositional expressions having the highest ones, and then tested on the most polarized cases of idioms and non-idioms.

For example, in the HtL setting, the AN bigrams we selected for the training set included idioms like *testa calda* 'hothead' and *faccia tosta* 'brazen person' (lit. 'tough face'), that reported an average idiomaticity rating of 6.8 and 6.6 out of 7 respectively, and non-idioms like *famoso scrittore* 'famous writer' and *nuovo governo* 'new government' that elicited an average idiomaticity rating of 1.2 and 1.1 out of 7. In the case of VN bigrams, we selected idioms like *andare a genio* 'to sit well' (lit. 'to go to genius') (mean idiomaticity rating of 7) and non-idioms like *vendere un libro* 'to sell a book' (mean idiomaticity rating of 1). The neural network was thus trained only on elements that our annotators had judged as clearly positive and clearly negative examples.

To provide examples on the LtH training sets, for the VN data, we selected idioms like *lasciare il campo* (mean rating = 3.6) and *cambiare colore* 'to change color (in face)' (mean rating = 3.6) against non-idiomatic expressions like *prendere un caffè* 'to grab a coffee' (3.3) and *lasciare un incarico* 'to leave a job' (2.3). For the AN data, we selected idioms like *prima serata* 'prime time' (lit. 'first evening') (mean rating = 4 out of 7) and compositional expressions like *proposta concreta* 'concrete proposal' (2.7). The neural network was in this case trained only on elements that our annotators had judged as borderline cases.

The results of these different filtering procedures can be found in Table 1.

7. Evaluation

Once the training sets were established, a variety of transformations were tried on our VN and AN distributional vectors before giving them as input to our network. Some models were trained on the raw 30,000 dimensional distributional vectors of VN and AN expressions; other models used the concatenation of the vectors of the individual components of the expressions; finally, other models employed PPMI (Positive Pointwise Mutual Information) (Church and Hanks 1991) and SVD (Singular Value Decomposition) transformed (Deerwester et al. 1990) vectors of 150 and 300 dimensions. Details of both classification and ordering tasks are shown in Table 1. Qualitative details about the results will be given in Section 8.

Table 1
Interpolated Average Precision (IAP), Spearman's ρ correlation with the human judgments and F-measure (F1) for Vector-Noun training (VN), Adjective-Noun training (AN), joint (VN+AN) training and training through vector concatenation. High-to-Low (HtL) models were trained on clear-cut cases, while Low-to-High (LtH) models were trained on borderline cases. As for the other models, the average performance over 5 runs with randomly selected training sets is reported. Training and test set are expressed as the sum of positive and negative examples.

Vectors	PPMI	SVD	Training	Test	IAP	ρ	F1
VN	Yes	No	15+15	30+30	.72	.48	.67
VN	Yes	No	20+20	25+25	**.73**	**.52**	**.77**
VN	Yes	150	15+15	30+30	.63	.35	.48
VN	Yes	150	20+20	25+25	.61	.33	.63
VN	Yes	300	15+15	30+30	.67	.33	.64
VN	Yes	300	20+20	25+25	.65	.3	.57
AN	No	No	8+8	6+4	.72	.19	.40
AN	Yes	No	8+8	6+4	.70	.06	**.60**
AN	Yes	150	8+8	6+4	.65	.11	.32
AN	Yes	300	8+8	6+4	**.88**	**.51**	.10
VN (HtL)	Yes	No	15+15	30+30	.71	.62	.77
VN (HtL)	Yes	No	20+20	25+25	**.79**	.65	.84
VN (LtH)	Yes	No	15+15	30+30	.71	.58	.80
VN (LtH)	Yes	No	20+20	25+25	.77	**.68**	**.85**
AN (HtL)	No	No	8+8	6+4	1	.8	.71
AN (HtL)	Yes	No	8+8	6+4	1	.71	.78
AN (LtH)	No	No	8+8	6+4	1	**.93**	**.89**
AN (LtH)	Yes	No	8+8	6+4	1	.84	.88
VN+AN	No	No	23+23	36+34 (joint)	.80	.64	.46
VN+AN (HtL)	No	No	23+23	36+34 (joint)	.63	.41	.65
VN+AN (LtH)	No	No	23+23	36+34 (joint)	**.68**	**.51**	**.66**
Conc. VN	No	No	20+20	24+24	.59	.34	.40
Conc. VN (HtL)	No	No	20+20	24 +24	**.61**	.07	.46
Conc. VN (LtH)	No	No	20+20	24+24	.57	**.31**	**.59**

7.1 Verb-Noun

We ran our model on the VN dataset, composed of 90 elements, namely 45 idioms and 45 non-idiomatic expressions. This is the largest of the two datasets. We trained our model on 30[3] and 40 elements for 20 epochs and tested it on the remaining 60 and 50 elements respectively. The models that best succeeded at classifying our phrases into idioms and non-idioms were trained with 40 PPMI-transformed vectors, reaching an average F1 score of .77 on the randomized iterations and an F1 score of .85, with a Spearman's ρ correlation of .68, when the training set was composed of borderline cases and the model was then tested on more clear-cut exemplars (LtH). As for the rest of the F1 scores,

3 When we report the number of training and test items in Table 1 as 15+15, for instance, we mean 15 idioms + 15 non-idioms. The same applies to all the other listed models.

what comes to light from our results is that increasing the number of training vectors generally leads to better results, except for models fed with SVD-transformed vectors of 300 dimensions, which seem to be insensitive to the size of our training data. Quite interestingly, SVD-reduced vectors appear to perform worse in general than raw ones and just PPMI-transformed ones. Due to space limitations, raw-frequency VN models are not reported in Table 1 since they were comparable to just PPMI-weighted ones.

This same pattern is encountered when evaluating the ability of our algorithm to rank idioms before non-idioms (IAP). The models with the highest score employs 40 PPMI training vectors and reach .73 on the randomized training, .79 on the HtL training and .77 on the LtH ones, while SVD training vectors generally lead to poorer ranking performances. Despite these IAP scores being encouraging, they are anyway lower than those obtained by Senaldi, Lebani, and Lenci (2016), who reach a maximum IAP of 0.91. This drop in performance could point to the fact that resorting to distributional information only to carry out idiom identification overlooks some aspects of the behavior of idiomatic constructions (e.g., formal rigidity) that is to be taken into account to arrive at a more satisfactory classification. Concerning the correlation between the continuous score of the neural net and the human idiomaticity ratings presented in Section 4.2, the best model also employed 40 PPMI vectors of borderline expressions (.68), followed by the model using 40 PPMI vectors of clear-cut cases (.65). These correlation values are quite comparable to the maximum of -0.67 obtained in Senaldi, Lebani, and Lenci (2016)[4] in High-to-Low and Low-to-High ordered models, while they are lower in randomized models, especially SVD-reduced ones.

All in all, both HtL and LtH experimental settings result in IAP, correlation and F1 scores that are higher than what we get from averaging over randomly selected training sets. More precisely, the strategy of training only on borderline examples (LtH) appears to be the most effective. This can intuitively make sense: once a network has learned to discriminate between borderline cases, detecting clear-cut elements should be relatively easy. The opposite strategy also seems to bring some benefits, possibly because training on clear negative and positive examples provides the network with a data set which is easier to generalize. In any case, it seems clear that selecting our training set with the help of human ratings allows us to significantly increase the performance of our models. We can see this as another proof that human continuous scores on idiomaticity - and not only binary judgments - are mirrored in the distributional pattern on these expressions. As for the influence of the training set size on IAP and ρ, all in all it seems that the best results are reached with 40 training vectors, both on the randomized training sets and on the ordered training sets.

The general trend we can abstract from these results is that our neural network does a good job in telling apart idioms and non-idioms by just relying on raw-frequency and PPMI-transformed distributional information. Performing dimensionality reduction apparently deprives the model of useful information, which makes the overall performance plummet to lower levels.

4 Please keep in mind that the correlation values in Senaldi, Lebani, and Lenci (2016) and Senaldi, Lebani, and Lenci (2017) are negative since the less similar a target vector to the vectors of its variants, the more idiomatic the target.

7.2 Adjective-Noun

Our model was also run on the AN dataset, composed of 26 elements (13 idioms and 13 non-idiomatic expressions). We empirically found that our network 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. As happened with VN vectors, performing SVD worsened the performance of the model. While F1 exact value can undergo fluctuations when a model is trained on very small sets, we always registered accuracies higher than 70% for the ordered training sets. In this case even more than in the Verb-Noun frame, the difference between randomizing the training set and selecting it using human idiomaticity ratings appears to be very evident, possibly due to the extremely small dimensions of this specific dataset, that make the qualitative selection of the training data of particular importance. Once again the highest Spearman's ρ correlation (.93) was reached when using a Low-to-High set trained on borderline cases, although it is important to keep in mind that such scores are computed on a very restricted test set. The same reasoning applies to IAP scores, which all reach the top value, though we must consider the very small test set. Senaldi, Lebani, and Lenci (2017) instead reached a maximum IAP of .85 and a maximum ρ of -.68 in AN idiom identification. 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 unseen 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.

7.3 Joint training

Our last experiment consisted in training our model on a mixed dataset of both VN and AN expressions, to check to what extent it would be able to recognize the same underlying semantic phenomenon across different syntactic constructions. In these models as well as in those described in Section 7.4, PPMI and SVD transformations were not tested anymore, since they were already shown to bring to generally comparable or even worse outcomes when tried on the VN and the AN datasets singularly. Concerning the structure of our training and test sets, two approaches were experimented with. We first tried to train our model on one pair type, e.g. the AN pairs, and then tested on the other, but we saw this required more epochs overall (more than 100) to stabilize and resulted in a poorer performance. When training our model on a mixed dataset containing the elements of both pair types, our model employed 20 epochs to reach an F-measure of 66% on the mixed training set when the set was ordered Low-to-High (i.e., it was composed of borderline cases only) and a comparable F-score of 65% when using clear-cut training input (HtL). Anyway, we also noticed that VN expressions were learned better than AN expressions. It's also worth considering that, although the F-scores of the LtH and HtL models were higher, the IAP and Spearman's ρ were lower than in the unordered input model. In other words, while ordering the input led to a better binary classification, the continuous scores returned a less precise ranking.

Our model was able to generalize over the two datasets, but this involved a loss in accuracy with respect to the only-VN and only-AN ordered training sets. It can be seen in Table 1 that a loss in accuracy is also evident for joint training on the randomized frame, although in this case the model seems hardly able to generalize at all.

7.4 Vector concatenation

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, Chatzikyriakidis, and Ghanimifard (2017). For example, instead of using the 30,000 dimensional vector of the expression *tagliare la corda*, we used the 60,000 dimensional vector resulting from the concatenation of *tagliare* and *corda*. This approach mimics the one adopted for metaphoric pairs and concludes our set of experiments, providing us with comparable results obtained from a compositionality-based approach to the same problem. 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 40 elements in training and 48 in testing and trained our model for 30 epochs overall. Predictably enough, vector composition resulted in the worst performance, differently from what happened with metaphors (Bizzoni, Chatzikyriakidis, and Ghanimifard 2017).

Despite all correlations are low and not statistically significant, it is still worth pointing out however that not all the results are completely random: with an F1 of 59% for the LtH training set and an IAP of .61 for the HtL set, 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, Glucksberg, and Cacciari 1994). Another hint in this direction is the difference in performance between randomized and ordered training that we can observe for concatenation: if human idiomaticity ratings were completely independent from the composition of the individual subparts of our idioms, such effect should not be present at all. Anyway, similarly to what happened with the joint models, ordering the training input led to higher F-scores and comparable IAPs, but returned a worse correlation with human judgments with respect to the models with a randomized training input.

8. Error Analysis

As we mentioned in Section 1, idiomaticity is not a black-or-white phenomenon and idioms are rather spread on a continuum of semantic transparency and formal rigidity, which makes some exemplars harder to classify. In our models we can find some "prototypical" cases of idioms which were always labeled correctly, like *toccare il fondo* 'to hit rock bottom', *lasciare il campo* and *passare alla storia* 'to go down in history' and also some cases of unambiguously classified non-idioms, like *andare in vacanza* 'to go on holiday', *ascoltare una canzone* 'to listen to a song' and *prendere un caffè*. On the other hand, we have some ambiguous expressions like *abbassare la guardia* 'to let down one's guard' and *sentire una voce* 'to hear a voice', which, despite being compositional and potentially literal, can be very often used figuratively, i.e. if someone were referring to *guardia* as a metaphorical defense or to *voce* as a rumor. In such cases, it might be the case that the evidence available in the chosen corpus privileged just one of the two possible readings, leading to labeling issues. By the same token, the expression *bussare alla porta (di qualcuno)* 'to go ask for (someone's) help' (lit.'to knock at the door'), which we initially labeled as idiomatic, can have a literal reading as well and that is why it was often labeled as non-idiomatic. Finally, as happened in Senaldi, Lebani, and Lenci (2016), some false positives like *chiedere le dimissioni* 'to demand the resignation' and *entrare in crisi* 'to get into a crisis' are compositional expressions which nonetheless display collocational behavior, since they represent very common and fixed expressions in the Italian language. Interestingly, while Senaldi, Lebani, and Lenci (2016) could

justify their being false positives since it is likely that the variant-based model took their lexical fixedness as a clue of their idiomatic status, our neural net relies on distributional semantic information only. What this suggests is that not only a semantic phenomenon like compositionality, but even a shallower one like collocability, which does not always and straightforwardly go hand in hand with non-compositionality, can be spotted out just by looking at contextual distribution.

As mentioned in Section 4.1, our target idioms and non-idioms varied considerably in frequency. We therefore conducted some correlation analyses to check out a possible relationship between the scores returned by our network and the frequency of our items. All in all, we can conclude that in most of our models frequency and the continuous idiomaticity scores were negatively correlated, though such a correlation did not show up systematically and was not always significant. In other words, the more frequent an item, be it an idiom or a literal, the more the network tended to consider it as literal (i.e., it gave it a lower idiomaticity score). This tendency could be explained if we consider that some of our most frequent idioms were actually quite ambiguous (e.g., *aprire gli occhi* 'to open one's eyes' occurred 6306 times in the corpus and *bussare alla porta* 3303 times) and most of their corpus occurrences could be literal uses.

9. Discussion and Conclusions

The experiments we have presented show 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. Specifically, we have observed that human continuous ratings of idiomaticity can be useful to select a better training set for our models, and that training our models on cases deemed by our annotators as borderline allows them to learn and perform better than if they were fed with randomized input. Also training our models only on clear-cut cases increases the performance. In general we can see from this phenomena that human continuous ratings of idiomaticity seem to be mirrored in the distributional structure of our data.

Unlike with metaphors (Bizzoni, Chatzikyriakidis, and Ghanimifard 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, like the ambiguity between idiomatic and literal meaning that some potentially idiomatic strings possess (e.g. *to leave the field*) and that would lead their contextual distribution to be more variegated with respect to only-literal combinations. We intend to further investigate this aspect in future works.

References

Abadi, Martín, 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*.

Artstein, Ron and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Baroni, Marco, 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.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 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, Baltimore, USA, June 22-27.

Bizzoni, Yuri, Stergios Chatzikyriakidis, and Mehdi Ghanimifard. 2017. "deep" learning: Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 7-11.

Blacoe, William and Mirella Lapata. 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, Jeju Island, Korea, July 12-14. Association for Computational Linguistics.

Bohrn, Isabel C., Ulrike Altmann, and Arthur M. Jacobs. 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, Cristina. 2014. Processing multiword idiomatic strings: Many words in one? *The Mental Lexicon*, 9(2):267–293.

Cacciari, Cristina and Costanza Papagno. 2012. Neuropsychological and neurophysiological correlates of idiom understanding: How many hemispheres are involved. *The handbook of the neuropsychology of language*, pages 368–385.

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

Collobert, Ronan and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167, Helsinki, Finland, July 5-9. ACM.

Cordeiro, Silvio, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 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, Berlin, Germany, August 7-12.

Cruse, D. Alan. 1986. *Lexical semantics*. Cambridge University Press.

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

Do Dinh, Erik-Lân and Iryna Gurevych. 2016. Token-level metaphor detection using neural networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*, pages 28–33, San Diego, USA, June 17.

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

Fazly, Afsaneh and Suzanne Stevenson. 2008. A distributional account of the semantics of multiword expressions. *Italian Journal of Linguistics*, 1(20):157–179.

Fraser, Bruce. 1970. Idioms within a transformational grammar. *Foundations of language*, pages 22–42.

Frege, Gottlob. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Geeraert, Kristina, R. Harald Baayen, and John Newman. 2017. Understanding idiomatic variation. In *Proceedings of the 13th Workshop on Multiword Expressions*, page 80, Valencia, Spain, April 4.

Gentner, Dedre. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Gibbs, Raymond W. 1993. Why idioms are not dead metaphors. *Idioms: Processing, structure, and interpretation*, pages 57–77.
Gibbs, Raymond W. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge University Press.
Gibbs, Raymond W., Josephine M. Bogdanovich, Jeffrey R. Sykes, and Dale J. Barr. 1997. Metaphor in idiom comprehension. *Journal of memory and language*, 37(2):141–154.
Glucksberg, Sam, Matthew S. McGlone, and Deanna Manfredi. 1997. Property attribution in metaphor comprehension. *Journal of memory and language*, 36(1):50–67.
He, Xinran and Yan Liu. 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, Cambridge, UK, February 6-10. ACM.
Klyueva, Natalia, Antoine Doucet, and Milan Straka. 2017. Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions*, page 60, Valencia, Spain, April 4.
Köper, Maximilian and Sabine Schulte im Walde. 2017. Improving verb metaphor detection by propagating abstractness to words, phrases and individual senses. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications*, pages 24–30, Valencia, Spain, April 4.
Krippendorff, Klaus. 2012. *Content analysis: An introduction to its methodology*. Sage.
Krčmář, Lubomír, Karel Ježek, and Pavel Pecina. 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, Sofia, Bulgaria, August 9.
Lakoff, George and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.
Legrand, Joël and Ronan Collobert. 2016. Phrase representations for multiword expressions. In *Proceedings of the 12th Workshop on Multiword Expressions*, Berlin, Germany, August 11.
Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Italian Journal of Linguistics*, 20(1):1–31.
Lenci, Alessandro. 2018. Distributional Models of Word Meaning. *Annual Review of Linguistics*, 4:151–171.
Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
Lin, Dekang. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, USA, June 20-26.
Liu, Dilin. 2003. The most frequently used spoken american english idioms: A corpus analysis and its implications. *Tesol Quarterly*, 37(4):671–700.
McGlone, Matthew S. 1996. Conceptual metaphors and figurative language interpretation: Food for thought? *Journal of memory and language*, 35(4):544–565.
McGlone, Matthew S., Sam Glucksberg, and Cristina Cacciari. 1994. Semantic productivity and idiom comprehension. *Discourse Processes*, 17(2):167–190.
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 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, Stateline, USA, December 5-10.
Mikolov, Tomas, 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, Altanta, USA, June 10-12.
Mitchell, Jeff and Mirella Lapata. 2010. Composition in Distributional Models of Semantics. *Cognitive Science*, 34(8):1388–1429.
Nunberg, Geoffrey, Ivan Sag, and Thomas Wasow. 1994. Idioms. *Language*, 70(3):491–538.
Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: Developing and Aligned Multilingual Database. In *Proceedings of the First International Conference on Global WordNet*, pages 293–302, Mysore, India, January 21-25.
Quartu, Monica B. 1993. *Dizionario dei modi di dire della lingua italiana*. RCS Libri.
Rei, Marek, Luana Bulat, Douwe Kiela, and Ekaterina Shutova. 2017. Grasping the finer point: A supervised similarity network for metaphor detection. *arXiv preprint arXiv:1709.00575*.

Rimell, Laura, Jean Maillard, Tamara Polajnar, and Stephen Clark. 2016. Relpron: A relative clause evaluation data set for compositional distributional semantics. *Computational Linguistics*, 42(4):661–701.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing*, pages 1–15, Mexico City, Mexico, February 17-23.

Senaldi, Marco S. G., Gianluca E. Lebani, and Alessandro Lenci. 2016. Lexical variability and compositionality: Investigating idiomaticity with distributional semantic models. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 21–31, Berlin, Germany, August 11.

Senaldi, Marco S. G., Gianluca E Lebani, and Alessandro Lenci. 2017. Determining the compositionality of noun-adjective pairs with lexical variants and distributional semantics. *Italian Journal of Computational Linguistics*, 3(1):43–58.

Steen, Gerard J., Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2014. A method for linguistic metaphor identification: From mip to mipvu. *Metaphor and the Social World*, 4(1):138–146.

Tanguy, Ludovic, Franck Sajous, Basilio Calderone, and Nabil Hathout. 2012. Authorship attribution: Using rich linguistic features when training data is scarce. In *PAN Lab at CLEF*, Valencia, Spain, September 23-26.

Titone, Debra and Maya Libben. 2014. Time-dependent effects of decomposability, familiarity and literal plausibility on idiom priming: A cross-modal priming investigation. *The Mental Lexicon*, 9(3):473–496.

Torre, Enrico. 2014. *The emergent patterns of Italian idioms: A dynamic-systems approach*. Ph.D. thesis, Lancaster University.

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

Wulff, Stefanie. 2008. *Rethinking Idiomaticity: A Usage-based Approach*. Continuum.