Probing for idiomaticity in vector space models

Marcos Garcia
CiTIUS Research Centre
Universidade de Santiago de Compostela
Galiza, Spain

Tiago Kramer Vieira
Federal University
of Rio Grande do Sul, Brazil

Carolina Scarton
University of Sheffield, UK

Marco Idiart
Federal University
of Rio Grande do Sul, Brazil

Aline Villavicencio
University of Sheffield, UK
Federal University
of Rio Grande do Sul, Brazil

marcos.garcia.gonzalez@udc.gal, tiagokv@hotmail.com,
c.scarton@sheffield.ac.uk, marco.idiart@gmail.com,
a.villavicencio@sheffield.ac.uk

Abstract

Contextualised word representation models have been successfully used for capturing different word usages, and they may be an attractive alternative for representing idiomaticity in language. In this paper, we propose probing measures to assess if some of the expected linguistic properties of noun compounds, especially those related to idiomatic meanings, and their dependence on context and sensitivity to lexical choice, are readily available in some standard and widely used representations. For that, we constructed the Noun Compound Senses Dataset, which contains noun compounds and their paraphrases, in context neutral and context informative naturalistic sentences, in two languages: English and Portuguese. Results obtained using four types of probing measures with models like ELMo, BERT and some of its variants, indicate that idiomaticity is not yet accurately represented by contextualised models.

1 Introduction

Contextualised word representation models, like BERT (Devlin et al., 2019) and ELMo (Peters et al., 2018), seem to represent words more accurately than static word embeddings like GloVe (Pennington et al., 2014), as they can encode different usages of a word. In fact, representations of a word in several contexts can be grouped in different clusters, which seem to be related to the various senses of the word (Schuster et al., 2019), and they can be used to match polysemous words in context to specific sense definitions (Chang and Chen, 2019). However, multiword expressions (MWEs) fall into a continuum of idiomaticity[1] (Sag et al., 2002; Fazly et al., 2009; King and Cook, 2017) and their meanings may not be directly related to the meanings of their individual words (e.g., *graduate student* vs. *eager beaver* as a hardworking person). Therefore, one question is whether and to what extent idiomaticity in MWEs is accurately incorporated by word representation models.

In this paper, we propose a set of probing measures to examine how accurately idiomaticity in MWEs, particularly in noun compounds (NCs), is captured in vector space models, focusing on some widely used representations. Inspired by the semantic priming paradigm (Neely et al., 1989), we have designed four probing tasks to analyse how these models deal with some of the properties of NCs, including non-compositionality (*big fish* as an important person), non-substitutability (*panda car* vs. *bear automobile*), or ambiguity (*bad apple* as either a rotten fruit or a troublemaker), as well as the influence of context in their representation. To do so, we have created the new Noun Compound Senses (NCS) dataset, containing a total of 9,220 sentences in English and Portuguese. This dataset includes sentence variants with (i) synonyms of the original NCs; (ii) artificial NCs built with synonyms of each component; or (iii) either the head or the modifier of the NC. Moreover, it is composed of naturalistic and controlled sense-neutral sentences, to minimise the possible effect of context words.

We compare five models (one static, GloVe, and four contextualised, ELMo and three BERT-based models) in English and Portuguese. The probing measures suggest that the standard and widely adopted composition operations display a limited ability to capture NC idiomaticity.

Our main contributions are: (i) the design of novel probes to assess the representation of id-

[1] We understand idiomaticity as *semantic opacity* and its continuum as different *degrees of opacity* (Cruse, 1986).

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 3551–3564
April 19 - 23, 2021.

iomaticity in vector models, (ii) a new dataset of NCs in two languages, and (iii) their application in a systematic evaluation of vector space models examining their ability to display behaviors linked to idiomaticity.

The remainder of this paper is organized as follows: First, Section 2 presents related work. Then, we describe the data and present the probing measures in Section 3. In Section 4, we discuss the results of our experiments. Finally, the conclusions of our study are drawn in Section 5.

2 Related Work

Priming paradigms have been traditionally used in psycholinguistics to examine how humans process language. For compounds, some findings suggest that idiomatic expressions are processed more slowly than semantically transparent ones, as processing the former may involve a conflict between the non-compositional and the compositional meanings (Gagné and Spalding, 2009; Ji et al., 2011). However, studies using event-related potential (ERP) data showed that idiomatic expressions, especially those with a salient meaning (Giora, 1999), have processing advantages (Laurent et al., 2006; Rommers et al., 2013). In NLP, probing tasks have been useful in revealing to what extent contextualised models are capable of learning different linguistic properties (Conneau et al., 2018). They allow for more controlled settings, removing obvious biases and potentially confounding factors from evaluations, and allowing both the use of artificially constructed but controlled sentences and naturally occurring sentences (Linzen et al., 2016; Gulordava et al., 2018). In priming tasks, related stimuli are easier to process than unrelated ones. One assumption is that, for models, related stimuli would achieve greater similarity than unrelated stimuli. These tasks have been used, for instance, to evaluate how neural language models represent syntax (van Schijndel and Linzen, 2018; Prasad et al., 2019), and the preferences that they may display, such as the use of mainly lexical information in a lexical substitution task even if contextual information is available (Aina et al., 2019).

Concerning pre-trained neural language models, which produce contextualised word representations, analyses about their abilities have shown, for instance, that they can encode syntactic information (Liu et al., 2019) including long-distance subject–verb agreement (Goldberg, 2019). Regarding semantic knowledge, the results of various experiments suggest that BERT can somewhat represent semantic roles (Ettinger, 2020). However, its improvements appear mainly in core roles that may be predicted from syntactic representations (Tenney et al., 2019). Moreover, from the representations generated by BERT, ELMo and Flair (Akbik et al., 2018) for word sense disambiguation, only the clusters of BERT vectors seem to be related to word senses (Wiedemann et al., 2019), although in cross-lingual alignment of ELMo embeddings, clusters of polysemous words related to different senses have also been observed (Schuster et al., 2019).

The use of contextualised models for representing MWEs has been reported with mixed results. Shwartz and Dagan (2019) evaluated different classifiers initialised with contextualised and non-contextualised embeddings in five tasks related to lexical composition (including the literality of NCs) and found that contextualised models, especially BERT, obtained better performance across all tasks. However, for capturing idiomaticity in MWEs, static models like *word2vec* (Mikolov et al., 2013) seem to have better performance than contextualised models (Nandakumar et al., 2019; King and Cook, 2018). These mixed results suggest that a controlled evaluation setup is needed to obtain comparable results across models and languages.

Therefore, we have carefully designed probing tasks to assess the representation of NCs in vector space models. As the same word can have different representations even in related paraphrased contexts (Shi et al., 2019), we adopt paraphrases with minimal modifications to compare the idiomatic and literal representations of a given NC.

3 Materials and Methods

3.1 Noun Compound Senses Dataset

The Noun Compound Senses (NCS) dataset is based on the NC Compositionality dataset, which contains NCs in English (Reddy et al., 2011), Portuguese and French (Cordeiro et al., 2019). Using the protocol by Reddy et al. (2011), human judgments were collected about the interpretation of each NC in 3 naturalistic corpus sentences. The task was to judge, for each NC, how literal the contributions of its component were for its meaning (e.g., "Is *climate change* truly/literally a *change* in *climate*?"). Each NC got a score, which was the average of the human judgments with a *Likert* scale from 0 (non-literal/idiomatic) to 5 (lit-

eral/compositional).[2]

For the NCS dataset, a set of probing sentences for the 280 NCs in English and the 180 NCs in Portuguese was added. For each NC, the sentences exemplify two conditions: (i) the naturalistic context provided by the original sentences (NAT), and (ii) a neutral context where the NCs appear in uninformative sentences (NEU). For the latter we use the pattern *This is a/an <NC>* (e.g., *This is an eager beaver*) and its Portuguese equivalent *Este/a é um(a) <NC>*. As some NCs may have both compositional and idiomatic meanings (e.g., *fish story* as either *an aquatic tale* or *a big lie*), these neutral contexts will be used to examine the representations that are generated for the NCs (and the sentences) in the absence of any contextual clues about the meaning of the NC. Moreover, they enable examining possible biases in the NC representation especially when compared to the representation generated for the NAT condition.

For each NC and condition, we created new sentence variants with lexical replacements, using synonyms of the NC as a whole or of each of its components. The synonyms of the NCs are the most frequent synonyms provided by the annotators of the original NC Compositionality dataset (e.g., *brain* for *grey matter*). The synonyms of each component were extracted from WordNet (Miller, 1995, for English) and from English and Portuguese dictionaries of synonyms (e.g., *alligator* for *crocodile* and *sobs* for *tears*). In cases of ambiguity (due to polysemy or homonymy), the most common meaning of each component was used. Experts (native or near-native speakers with linguistics background) reviewed these new utterances, keeping them as faithful as possible to the original ones, but with small modifications for preserving grammaticality after the substitution (e.g., modifications in determiners and adjectives related to gender, number and definiteness agreement).

NCS contains a total of 5,620 test items for English and 3,600 for Portuguese among neutral and naturalistic sentences, and it is freely available.[3]

[2]We averaged the Likert judgments for comparability with previous work, even though the median may reflect better the cases where there is more disagreement among the annotators. However, both mean and median are strongly correlated in our data: $\rho = 0.98$ (English) and $\rho = 0.96$ (Portuguese), $p < 0.001$.

[3]https://github.com/marcospln/noun_compound_senses

3.2 Probing Measures

This section presents the probing measures defined to assess how accurately idiomaticity is captured in vector space models. For these measures we consider comparisons between three types of embeddings: (i) the embedding for an NC out of context (i.e. the embedding calculated from the NC words alone), represented by ϵ_{NC}; (ii) the embedding for an NC in the context of a sentence S, represented by $\epsilon_{\text{NC} \subset \text{S}}$[4] (iii) finally, the sentence embedding that contains an NC, which is represented by $\epsilon_{\text{S} \supset \text{NC}}$. Here we use the standard output of some widely used models with no fine-tuning to avoid possible interference. However, in principle, these measures could apply to any embedding even after fine-tuning.

The similarities between embeddings are calculated in terms of cosine similarity: $\cos(\epsilon, \epsilon')$ where ϵ and ϵ' are embeddings from the same model with the same number of dimensions. In NAT cases, the similarity scores for each of the three available sentences for a given NC are averaged to generate a single score. We use Spearman ρ correlation between similarities and the NC idiomaticity scores (280 for English and 180 for Portuguese) to check for any effects of idiomaticity in the probing measures. We also calculate Spearman ρ correlation between different embedding models to determine how much the models agree, and between the NAT and NEU conditions to see how much the context affects the distribution of similarities. We also analyse the distribution of cosine similarities produced by different models for each of the probing measures. All probing measures are calculated for both NAT and NEU conditions.

P1: Probing the similarity between an NC and its synonym. If a contextualised model captures idiomaticity accurately, the embedding for a sentence containing an NC should be similar to the embedding for the same sentence containing a synonym of the NC (NC_{syn}, e.g., for *grey matter*, NC_{syn} = *brain*). Thus, $\text{sim}^{(\text{P1})}_{\text{Sent}} \simeq 1$, where $\text{sim}^{(\text{P1})}_{\text{Sent}} = \cos(\epsilon_{\text{S} \supset \text{NC}}, \epsilon_{\text{S} \supset \text{NC}_{syn}})$. This should occur regardless of how idiomatic the NC is, that is, similarity scores are not expected to correlate with NC idiomaticity scores ($\rho^{(\text{P1})}_{\text{Sent}} \simeq 0$). Moreover, this should also hold for the NC and NC_{syn} embeddings generated in the context of this sentence, which means that $\rho^{(\text{P1})}_{\text{NC}} \simeq 0$ and $\text{sim}^{(\text{P1})}_{\text{NC}} \simeq 1$

[4]For non-contextualised embeddings $\epsilon_{\text{NC} \subset \text{S}} = \epsilon_{\text{NC}}$.