Don't Invite BERT to Drink a Bottle: Modeling the Interpretation of Metonymies Using BERT and Distributional Representations

Paolo Pedinotti **Alessandro Lenci**
CoLing Lab, University of Pisa, Italy
pedinotti.paolo@gmail.com alessandro.lenci@unipi.it

Abstract

In this work, we carry out two experiments in order to assess the ability of BERT to capture the meaning shift associated with metonymic expressions. We test the model on a new dataset that is representative of the most common types of metonymy. We compare BERT with the Structured Distributional Model (SDM), a model for the representation of words in context which is based on the notion of Generalized Event Knowledge. The results reveal that, while BERT ability to deal with metonymy is quite limited, SDM is good at predicting the meaning of metonymic expressions, providing support for an account of metonymy based on event knowledge.

1 Introduction

Metonymy is one of the most important sources of lexical polysemy and consists in the **meaning shift** of a noun that is used to refer to another entity to which it is related (Littlemore, 2015). For instance, *bottle* refers to a solid container in (1a), but in (1b) it stands for some liquid contained in it:

(1) a. The guest broke the **bottle**.
b. The guest tasted the **bottle**.

Metonymy is a productive and systematic process (e.g., all nouns denoting containers show the same polysemy as *bottle*, giving rise to the so-called CONTAINER-FOR-CONTENT metonymic alternation). Therefore, both linguistic (Pustejovsky, 1995; Jackendoff, 1997; Asher, 2011) and psycholinguistic (Piñango et al., 2016) studies contest the treatment of metonymy like a case of lexical ambiguity, and instead support the hypothesis that metonymic interpretations result from the inherently dynamic and generative nature of lexical representations that can acquire new meanings by integrating information activated by the textual and extralinguistic context.

Vector representations (aka *word embeddings*) produced by Distributional Semantic Models (DSMs) are particularly suitable for modeling contextual semantic effects, due to their "gradedness" and their dependence on the linguistic contexts (Lenci, 2018; Boleda, 2020). Traditional DSMs represent the content of lexical types through a single vector that "summarizes" their whole distributional history. Things have recently changed with the introduction of deep neural architectures for language modeling like BERT (Devlin et al., 2019), whose word representations have helped achieving state-of-the-art results in a wide variety of supervised NLP tasks. These embeddings are intrinsically *contextualized*, in the sense that the model computes a different vector for each token occurrence of the same word, depending on the sentence in which the token appears. In this work, we test whether BERT contextualized embeddings can be used to model the meaning shifts associated with metonymic uses of words. Given its pervasiveness in everyday communication, we suggest that the extent to which metonymy is captured by BERT is an important testbed to evaluate its actual ability to model natural language. In line with this goal, we require the model to induce the additional meaning of metonymic expressions by encoding it into the contextualized embeddings. We also compare BERT performance with that achieved by the Structured Distributional Model by Chersoni et al. (2019), in which the context-sensitive nature of lexical meaning

Proceedings of the 28th International Conference on Computational Linguistics, pages 6831–6837
Barcelona, Spain (Online), December 8-13, 2020

is instead captured by integrating a rich array of distributional knowledge about events and their typical participants (McRae and Matsuki, 2009).

2 Related Work

Over the years, various methods to obtain contextualized representations of word meaning have been developed in different fields. Research in distributional semantics (Erk and Padó, 2008; Thater et al., 2011) has taken non-contextual representations of words as starting point from which contextualized vectors capable of modeling various types of meaning alternations are derived (Erk and Padó, 2008; Zarcone et al., 2012). However, these models have never been used to predict metonymic semantic shifts. Lately, Transformer language models (e.g., BERT, GPT2, etc.) have stormed AI and NLP with a new generation of word embeddings that are expected to capture lexical meaning variation in context (Radford et al., 2019; Devlin et al., 2019). In particular, the representations produced by BERT (Devlin et al., 2019) have been used to create high performing models for many language understanding tasks, although their status as a linguistically sound model of meaning is debated (Mickus et al., 2020). Shwartz and Dagan (2019) test BERT on several cases of figurative language, but to the best of our knowledge BERT ability to identify metonymy has never been addressed yet.

3 Models

In BERT, the embedding of a word is modified with contextual information through the self-attention mechanism of Transformers (Vaswani et al., 2017). As is well known, BERT is trained on two tasks: predicting randomly masked tokens (Masked Language Model) and determining whether a sentence follows another sentence in a dataset (Next Sentence Prediction). Since our intent is to assess the model ability to understand metonymic meanings, we test the contextual embeddings themselves (**BERT-Emb**), rather than fine-tuning them in a supervised classification task (cf. Mickus et al., 2020 for a similar approach). We also investigate if this ability is reflected in the probabilities that the model assigns to the masked metonymic word, thereby exploiting BERT as a language model (**BERT-LM**).

The model against which we evaluate BERT is inspired by the Structured Distributional Model by Chersoni et al. (2019), which is based on the notion of Generalized Event Knowledge (GEK; McRae and Matsuki, 2009). GEK is conceptual knowledge about real-world events and their participants, which has been shown to influence sentence processing by causing expectations regarding the upcoming input. For example, when the first part of a sentence starting with the words *The police arrested* is processed, expectations about the possible objects are generated based on knowledge about the typical patients of the event *arrest* (e.g., *thief*, *burglar*, etc.). Since GEK becomes activated quickly, it has been argued that the meaning assigned to words in context comes as a result of the interaction between lexical meaning and the expectations generated (Elman, 2014).

The model presented here uses the graph-based distributional model of event knowledge introduced in Chersoni et al. (2019) to compute a contextualized representation of word meaning, which is obtained by integrating the lexical embedding of a word with a vector representation of the expectations activated by the context for the event role of the word. As in the model of Erk and Padó (2008), we approximate the *expectations* activated by a word *w* by selecting the words with the highest pointwise mutual information (PMI) with *w* in a corpus.[1] We use a parsed corpus to extract different expectations according to their syntactic roles, as a surface approximation of semantic roles (e.g., typical patients are derived from the verb direct objects). For example, given the sentence *The guest tasted the bottle*, we consider the most typical objects of *w* (the verb *taste*), which provide an approximation of the typical patients of the event expressed by the word. We take the direct objects since this is the function of the metonymic word w_m in the sentence (i.e., the noun *bottle*). A key feature of the model is that the activated words are filtered according to their PMI association strength with the metonymic word w_m. In our example, words for foods (e.g., *fruit*, *meat*, etc.) and drinks (e.g., *wine*, *beer*, etc.) are activated by *taste*, but the former are discarded because they have low PMI values with *bottle*. This process, which is an original innovation

[1]We take only the expectations activated by the verb in a direct syntactic relation to the metonymic word, even though proposals to integrate expectations from all the sentence arguments have been developed (Lenci, 2011; Chersoni et al., 2019)

with respect to the Chersoni et al. model, simulates the interaction between lexical information and active expectations as described by Elman (2014), and at the same time reproduces the associative processes involved in metonymy interpretation by updating the salience of expectations based on their relation with the metonymic word. Finally, we calculate the centroid of the activated expectation vectors and the embedding of w_m to obtain its contextualized representation.

Let W be the *k* words with the highest PMI with the verb *w* and W_m the *n* words in W with the highest PMI with the metonymic word w_m (for all experiments, we set *k*=30 and *n*=5). The contextualized representation of the metonymic word $\overrightarrow{w'_m}$ is built by summing the lexical vectors of the words in W_m and the metonymic word w_m. Each element of the resulting vector is then divided by the number of words used for the creation of the vector (equivalent to *n*+1). This procedure extends the notion of mean to a vector space to produce context-adapted representations of word meaning which are comparable with lexical vectors. Finally, $\overrightarrow{w'_m}$ is defined as follows:

$$\overrightarrow{w'_m} = \frac{((\sum_{w \in W_m} \overrightarrow{w}) + \overrightarrow{w_m})}{|W_m| + 1} \quad (1)$$

4 Dataset and experiments

4.1 Dataset

We introduce a new dataset that is representative of the most common types of metonymy, which we make available to the research community.[2] The dataset includes 509 items, each consisting of two sentences: i.) a sentence where a target word (e.g., *bottle*) is used metonymically (e.g., *The guest tasted the bottle*), together with a paraphrase making explicit the metonymic meaning (metonymic paraphrase, e.g., *wine*), and ii.) a sentence where the same word occurs with its literal meaning (e.g., *The man raised the bottle*), together with a paraphrase making it explicit (literal paraphrase, e.g., *container*). The metonymy types represented in the dataset are: CONTAINER-FOR-CONTENT (*The guest tasted the **bottle** → wine*) (Radden and Kövecses, 1999), PRODUCER-FOR-PRODUCT (*The **author** is translated into the language → novels*) (Radden and Kövecses, 1999), PRODUCT-FOR-PRODUCER (*The **newspaper** hates the politician → editor*) (Handl, 2011), LOCATION-FOR-LOCATED (*The **theater** applauded the performers→ audience*) (Barcelona, 2015), CAUSER-FOR-RESULT (*The **fans** than drowned out the announcer → screams*) (Warren, 2006), POSSESSED-FOR-POSSESSOR (*76 **trombones** marched into the park → musicians*) (Radden and Kövecses, 1999).

4.2 Experiments

We perform two different experiments to determine whether the models reproduce the whole set of semantic relations described in each item of the dataset. There are two different versions of each experiment. The first version is designed to be carried out using contextualized embeddings produced by BERT and SDM, the second using BERT as a language model.

Experiment 1 – The goal is to verify whether a model is able to detect the meaning shift associated with metonymy by representing the new meaning at the same time. **Contextualized embeddings (BERT, SDM)**: We test whether the similarity relations described in the dataset between a word and its paraphrase are reproduced in the structure of the vector spaces produced by the models. We can infer from the items of the dataset the following structure of semantic relations: the contextual meaning resulting from the metonymic usage of a word (e.g., the meaning of *bottle* in *The guest tasted the bottle*) is more similar to the meaning of a possible metonymic interpretation (e.g., *wine*) and less similar to the meaning of the same word used in its literal sense (e.g., in *The man raised the bottle*). For each test item, we feed the models with the metonymic sentence (e.g., *The guest tasted the bottle*) and we take the model representation of the target word ($\overrightarrow{met}$). Then, we feed the models with the literal sentence (e.g., *The man raised the bottle*) and we take the model representation of the target word ($\overrightarrow{lit}$). Finally, we feed the models with the metonymic sentence in which the target word has been replaced with the metonymic para-

[2]The dataset is available here: `https://github.com/ppedin/MetonymyData`. Sentences were manually extracted from online corpora of written English from various genres.

phrase (e.g., *The guest tasted the wine*). This time, we take the model representation of the metonymic paraphrase ($\overrightarrow{metpar}$), which we use as a ground-truth representation of the metonymic meaning. We expect the model to satisfy the inequality $sim(\overrightarrow{met}, \overrightarrow{metpar}) > sim(\overrightarrow{met}, \overrightarrow{lit})$, if metonymy is interpreted correctly (sim = cosine similarity). **Language Model (BERT)**: We examine whether, given the surrounding context of a word that receives a metonymic interpretation (like the sequence *The guest tasted*), the model is able to compute a representation of the most plausible completions of the context that match data from the dataset, namely that the corresponding metonymic sense (e.g., *wine*) is preferred to the literal interpretation of a word like *bottle*. For each test item, we feed BERT with the metonymic sentence in which the target word has been masked (e.g., *The guest tasted the* [MASK]). We then get the probabilities of the target word (e.g., *bottle*) and its metonymic paraphrase (e.g., *wine*). If the preference for the metonymic interpretation is reflected in BERT prediction, then the probability of the metonymic paraphrase is expected to be higher than that of the target.

Experiment 2 – The goal is to test the model ability to associate each target word occurrence with the corresponding (literal vs. metonymic) sense. The experiment consists of two subtasks: **Metonymic Matching** and **Literal Matching**. **Contextualized embeddings (BERT, SDM)**: We follow the same methodology from Experiment 1, but this time we use a more extensive set of semantic similarity relations from the dataset. We consider the following relations: Compared to the literal usage of the same word (e.g., *The man raised the bottle*), the semantic representation for the metonymic usage of a word (e.g., *The guest tasted the bottle*) is more similar to a possible metonymic interpretation (e.g., *wine*), and at the same time is less similar to a paraphrase of its literal meaning (e.g., *container*). We create two new sentences for each test item, one with the metonymic paraphrase (e.g., *The wine steward decanted the wine*), and another one with the literal paraphrase (e.g., *The customer fills the container*) so that we can extract contextualized representations of the paraphrases which are directly comparable with those of the target word. As in the previous experiment, we feed the models with the metonymic (e.g., *The guest tasted the bottle*) and the literal sentence (e.g., *The man raised the bottle*) and we take the representations of the target word ($\overrightarrow{met}$ and $\overrightarrow{lit}$ respectively). Then, we feed the models with the newly created sentence with the metonymic paraphrase (e.g., *The wine steward decanted the wine*) and we take the model representation of the paraphrase ($\overrightarrow{metpar}$), which we use as a ground-truth representation of the metonymic sense. Finally, we feed the models with the newly created sentence with the literal paraphrase (e.g., *The customer fills the container*) and we take the model representation of the paraphrase ($\overrightarrow{litpar}$), which we use as a ground-truth representation of the literal meaning of the target word. In the Metonymic Matching subtask, we assess whether the models satisfy the inequality $sim(\overrightarrow{met}, \overrightarrow{metpar}) > sim(\overrightarrow{lit}, \overrightarrow{metpar})$. In the Literal Matching subtask, we assess whether the models satisfy the condition $sim(\overrightarrow{lit}, \overrightarrow{litpar}) > sim(\overrightarrow{met}, \overrightarrow{litpar})$. **Language Model (BERT)**: We adopt the same methodology used for Experiment 1. We use BERT language model to compute a representation of the most likely completions of the surrounding context of a metonymic word (like the sequence *The guest tasted*). We examine whether the representation reflects the fact that a possible metonymic sense of a word like *bottle* (e.g., *wine*) is preferred to the literal interpretation of the word. This time, we use a paraphrase of the literal meaning of the target word (e.g., *container*) instead of the word itself. Moreover, we do the same for the context in which the word occurs with its literal sense (e.g., the sequence *The man raised*) and we investigate whether the representation expresses the preference for the literal meaning. For each test item, we feed BERT with the metonymic and the literal sentences with the target word masked (*The guest tasted the* [MASK] and *The man raised the* [MASK] respectively). We then compare the probabilities of the metonymic (e.g., *wine*) and the literal paraphrase (e.g., *container*). We expect the former to be higher than the latter in the metonymic sentence (Metonymic Matching subtask), and the opposite to be true in the literal sentence (Literal Matching subtask).

We use $\text{BERT}_{\text{BASE}}$ (number of layers=12, hidden size=768, number of self-attention heads=12) in all experiments. To implement SDM, we produce 300-dimensional dependency-based embeddings using Skip-gram with negative sampling (Levy and Goldberg, 2014) trained on a parsed corpus of about 3.9 billion tokens, which is a concatenation of ukWaC and a 2018 dump of Wikipedia.

	Experiment 1			Experiment 2					
				Metonymic Matching			Literal Matching		
	BERT		**SDM**	**BERT**		**SDM**	**BERT**		**SDM**
Type of Metonymy (#Items)	**Emb**	**LM**		**Emb**	**LM**		**Emb**	**LM**	
CONTAINER-FOR-CONTENT (89)	0.37	0.53	0.78	0.56	0.76	0.71	0.57	0.45	0.66
PRODUCER-FOR-PRODUCT (110)	0.59	0.80	0.95	0.63	0.49	0.81	0.71	0.86	0.59
PRODUCT-FOR-PRODUCER (47)	0.47	0.23	0.96	0.70	0.53	0.62	0.62	0.91	0.75
LOCATION-FOR-LOCATED (94)	0.39	0.52	0.82	0.66	0.52	0.75	0.78	0.89	0.80
CAUSER-FOR-RESULT (92)	0.17	0.71	0.84	0.72	0.79	0.69	0.70	0.78	0.66
POSSESSED-FOR-POSSESSOR (77)	0.45	0.55	0.86	0.60	0.73	0.62	0.69	0.61	0.65
All	0.41	0.59	**0.87**	0.64	0.64	**0.72**	0.68	**0.75**	0.69

Table 1: Accuracy of the models in the two experiments.

5 Results and discussion

The results of the experiments are presented in Table 1. We report model accuracy in satisfying the expected inequality conditions for each subtask.

In Experiment 1, SDM largely outperforms BERT in all metonymy types. These results indicate that SDM is particularly effective in deriving the additional meaning (the average cosine similarity between the representation of a metonymic word and its paraphrase is 0.79). On the other hand, BERT ability to deal with metonymy is much more limited, although performance varies considerably both within and between the two methods we used. This variability is interesting, as it suggests that the various metonymy types have different properties that deserve more in-depth analysis and might call for different computational solutions. However, BERT generally achieves higher accuracy when used as a language model (0.59 vs. 0.41). This can be attributed to the fact that BERT-LM is based on information that is similar to that used by SDM (i.e., context-based predictions). However, SDM also explicitly integrate word meaning with general world knowledge in the form of typical event participants, which could explain its better performance.

The results of Experiment 2 indicate that BERT is much more accurate when asked to choose between two possible interpretations (metonymic and literal) of the same word. However, the results can be viewed as supporting the findings of the first experiment since i.) SDM performance on the Metonymic Matching subtask (involving the association of metonymic words with their interpretations) is generally higher than the other methods, and ii.) while SDM performs better on the former subtask than on the latter, the opposite is true for BERT. Again, important variability between metonymy types and between methods can be observed. In particular, BERT-LM scores for some metonymy types on the Literal Matching subtask are significantly high, confirming the trend that BERT generally produces more accurate predictions about the interpretation of words in context when used as a language model.

6 Conclusion

We have shown that BERT effectiveness in modeling word meaning in context is quite limited when a metonymic shift is involved. On the other hand, a model like SDM that simulates the associative aspects of sentence processing potentially involved in metonymy produces contextualized representations that encode a significant amount of information about metonymic meaning. These results have potential implications for linguistic theory, since they suggest a relationship between metonymic meaning and the conceptual event-based expectations that we produce during processing, contributing to a psycholinguistic model of how metonymy (and language in general) is interpreted. This finding is further corroborated by the fact that BERT yields better results when its predictions about the interpretation of metonymic expressions are based on the probabilities it assigns to words when used as a language model.

References

Nicholas Asher. 2011. *Lexical Meaning in Context*. Cambridge University Press, Cambridge, UK.

Antonio Barcelona. 2015. Metonymy. In Ewa Dąbrowska and Dagmar Divjak, editors, *Handbook of Cognitive Linguistics*, page 143–167. De Gruyter Mouton, Berlin/Boston.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6:213–23.

Emmanuele Chersoni, Enrico Santus, Ludovica Pannitto, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2019. A structured distributional model of sentence meaning and processing. *Natural Language Engineering*, 25:483–502.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 4171–4186, Minneapolis, Minnesota.

Jeffrey L. Elman. 2014. Systematicity in the Lexicon: On Having your Cake and Eating it too. In Paco Calvo and John Symons, editors, *The Architecture of Cognition: Rethinking Fodor and Pylyshyn's Systematicity Challenge*, pages 1–33. The MIT Press, Cambridge, MA.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, page 897–906, Honolulu, Hawaii.

Sandra Handl. 2011. *The Conventionality of Figurative Language: A Usage-based Study*. BoD – Books on Demand, Norderstedt, Germany.

Ray Jackendoff. 1997. *The architecture of the Language Faculty*. MIT Press, Cambridge, MA.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pages 58–66, Portland, OR.

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

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, page 302–308, Baltimore, Maryland.

Jeannette Littlemore. 2015. *Metonymy: Hidden Shortcuts in Language, Thought and Communication*. Cambridge University Press, Cambridge, UK.

Ken McRae and Kazunaga Matsuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Lang Linguist Compass*, 3(6):1417–1429.

Timothee Mickus, Denis Paperno, Mathieu Constant, and Kees van Deemter. 2020. What do you mean, bert? assessing bert as a distributional semantics model. *Proceedings of the Society for Computation in Linguistics*, 3.

Maria M. Piñango, Muye Zhang, Emily Foster-Hanson, Michiro Negishi, Cheryl Lacadie, and R. Todd Constable. 2016. Metonymy as referential dependency: Psycholinguistic and neurolinguistic arguments for a unified linguistic treatment. *Cognitive Science*, 41.

James Pustejovsky. 1995. *The Generative Lexicon*. MIT Press, Cambridge, MA.

Günter Radden and Zoltán Kövecses. 1999. Towards a Theory of Metonymy. In Klaus-Uwe Panther and Günter Radden, editors, *Metonymy in Language and Thought*, pages 17–59. John Benjamins, Amsterdam/Philadelphia.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, and Dario Amodei abd Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word meaning in context: A simple and effective vector model. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, page 1134–1143, Long Beach, CA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, Long Beach, CA.

Beatrice Warren. 2006. Referential metonymy. *Scripta Minora Regiae Societatis Humaniorum Litterarum Lundensis*.

Alessandra Zarcone, Jason Utt, and Sebastian Padó. 2012. Modeling covert event retrieval in logical metonymy: probabilistic and distributional accounts. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 70–79, Montréal, Canada.