Interpreting Logical Metonymy through Dense Paraphrasing

Bingyang Ye* **(byye@brandeis.edu)**
Brandeis University, Department of Computer Science, 415 South Street, Waltham, MA 02453 USA

Jingxuan Tu* **(jxtu@brandeis.edu)**
Brandeis University, Department of Computer Science, 415 South Street, Waltham, MA 02453 USA

Elisabetta Jezek (jezek@unipv.it)
University of Pavia, Department of Humanities, Strada Nuova 65, 27100 Pavia, Italy

James Pustejovsky (jamesp@brandeis.edu)
Brandeis University, Department of Computer Science, 415 South Street, Waltham, MA 02453 USA

Abstract

Compositionality has been argued to be both a desirable and perhaps even necessary component of interpreting language, yet there appear to be many linguistic phenomena that do not overtly exhibit semantic compositional behavior. One of the most interesting challenges involves the phenomena of contextual modulations referred to collectively as semantic coercion or logical metonymy. Some models of how we understand, for example, *Alex enjoyed her coffee* and *Jen heard the train* incorporate mechanisms that provide "compositional flexibility", to allow event-selecting and sound-selecting verbs, respectively, to combine with arguments that denote neither. In this paper, we present a computational model that provides for such flexibility in the interpretation of a verb with its arguments, for such coercive contexts in English. Specifically, we argue that such constructions typically have surface structural correlates in the form of *dense paraphrases*, and that these forms can be used to model the masked content in the coerced compositional context. We present preliminary results using a transformer architecture (BERT) on a masked completion task. This suggests that constructions involving "enriched composition" can in fact be computationally analyzed with attention-based architectures. Our results show that modeling logical metonymy is a challenging task but can be substantially improved by fine-tuning through dense paraphrasing.

Keywords: semantic coercion; logical metonymy; compositionality; polysemy; transformers; distributional semantics

Introduction

The question of how functional expressions such as verbs and prepositions impose semantic constraints on their arguments has long been one of the major research themes in theoretical linguistics (Katz & Fodor, 1963; Chomsky, 1965; Lakoff, 1970; Jackendoff, 1972), as well as in formal treatments of type-driven selection (Partee, 1973; Dowty, 1979), and cognitive approaches to frames (Fillmore, 2008). Within these traditions, two types of predicative selection on an argument can be distinguished: (a) thematic role or semantic relation identification (AGENT, THEME, PATIENT, etc.); and semantic type selection on the argument (EVENT, PHYSOBJ, PROPOSITION, etc.). While the former addresses how a verb's arguments participate in the frame or situation denoted by an event (Van Valin, 1999; Dowty, 1991), we focus here on the second issue, that of *type selection*: what semantic types are imposed (or selected) by a verb on its arguments, given a specific verb sense (Jackendoff, 1990; Pollard & Sag, 1994). We examine the selection mechanisms involved in logical metonymy in language (Apresjan, 1974; Pustejovsky, 1995; Asher, 2011) and how they can be computationally interpreted using recent transformer architectures, such as BERT (Devlin, Chang, Lee, & Toutanova, 2019). In particular, we propose that metonymic constructions typically have surface structural correlates in the form of *dense paraphrases*, and that these forms can be used to model the masked content in the coerced compositional context. This suggests that constructions involving "enriched composition" (Pustejovsky, 1995; Jackendoff, 1997) can in fact be computationally analyzed with attention-based architectures.

Cases of logical metonymy via enriched composition involve constructions where the type expected by a predicate is not what is superficially present in the argument. Consider the range of verb-object selections illustrated in (1-3). Some predicates seem to directly select their argument, as in (1), where the type expected (selected for) by the verb is directly matched by the direct object's type.

(1) a. The dog **ate** [the biscuit]$_{\text{FOOD}}$.
b. The girl **heard** [a sound]$_{\text{SOUND}}$.
c. The Senator **believes** [she is innocent]$_{\text{PROPOSITION}}$.

Now consider examples where the selection seems less direct, involving the construal of "missing material", marked in brackets, [. . .].

(2) a. The dog **enjoyed** [*eating*] the biscuit.
b. The girl **heard** [*the sound of*] a dog.
c. Jen **tied** [*the laces of*] her shoes.

In each of these sentences, the type selected by the verb is satisfied by a kind of semantic reconstruction. In fact, the construal in (2a) is common with both sentiment and aspectual predicates more broadly (Pustejovsky & Bouillon, 1995).

(3) a. Alex **finished** [*writing/reading*] the letter.
b. The chorus **began** [*singing*] the song.

[*] These authors contributed equally to this work

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

The material in brackets, [...], is not part of the surface form of the sentence, but is a possible grammatical construal of the missing verb.

In the remainder of the paper, we show how semantic treatments of logical metonymy can be interpreted as dense paraphrasing models using attention-based architectures such as BERT. The claim is that logical metonymy, and coercive contexts more generally, will usually reveal a set of paraphrases that act as the signature for a semantic type. This is in fact consistent with Harris's (Harris, 1970) original notion of transformational set, as well as the *canonical syntactic forms* for a semantic type (Pustejovsky, 1995).

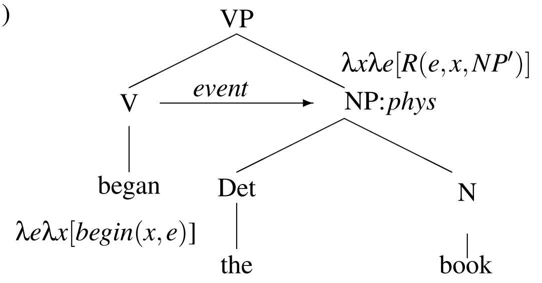
Approaches to Logical Metonymy

Large linguistic corpora present many challenges to the notion that language is compositional (Melamed, 1997; Holsinger, 2009; Bu, Zhu, & Li, 2010; Pustejovsky, 2012; Yazdani, Farahmand, & Henderson, 2015). In these studies, we distinguish between the non-compositionality of idioms and multiword expressions (MWEs) on the one hand, and focus in this paper on deeper interpretive contexts such as logical metonymy.

Interpreting logical metonymy from the underlying semantic level involves a type-change, such as coercion (Pustejovsky & Jezek, 2008; Asher, 2011), where the type is reconstructed in the composition of the surface expression. For example, the logical metonymy of the NP "the book" to the type EVENT in (4b) is computed by coercion shown in (5).

(4) a. Mary began [reading the book]$_{\text{EVENT}}$.
b. Mary began [the book]$_{\text{EVENT}}$.

(5)


VP
$\lambda x \lambda e[R(e,x,NP')]$
V
event
NP:*phys*
began
Det
N
$\lambda e \lambda x[begin(x,e)]$
the
book

In such configurations, the verb is said to "coerce" the NP argument into an event interpretation. Under such an analysis, the NP may denote a salient event that involves the book in some way, e.g., reading it, writing it, and so on, perhaps part of the Qualia Structure of the head. This is schematically represented above, where the NP *the book* has been reinterpreted through coercion, as embedded within a relation, R, and a subsequent event, e, involving the book.

Alternatively, it is possible to view logical metonymy as a relation between surface forms in the language. This was, in fact, proposed by (Harris, 1970), where such verb-complement constructions admitted of multiple paraphrases, as seen in (4) above. While the coercion model assumes an underlying semantic type even in the absence of observable signatures matching that type, the paraphrase model can be seen as learning alternative structural forms, and discovering a paraphrase set that can act as the signature to a semantic type. In fact, we argue that these two positions are complementary views on a richer model of compositionality.

Harris' view and that presented in (Smaby, 1971) is related to recent attempts to enrich surface sentence forms that are missing information through "decontextualization" procedures that textually supply information which would make the sentence interpretable out of its local context (Choi et al., 2021; Elazar, Basmov, Goldberg, & Tsarfaty, 2021; Wu, Luan, Rashkin, Reitter, & Tomar, 2021). Paraphrasing and decontextualizing are closely related, and in fact part of a richer process of what we call *dense paraphrasing* (Self, n.d.). This combines the textual variability of an expression's meaning (paraphrase) with the amplification or enrichment of meaning associated with an expression (decontextualization).

While a paraphrase is typically defined as a relation between two expressions that convey the same meaning (Bhagat & Hovy, 2013), it has also been used to clarify meaning through verbal, nominal, or structural restatements that preserve (and enhance) meaning (Smaby, 1971; Kahane, 1984; Mel'čuk, 1995, 2012), in particular the notion of "entailed paraphrase" (Culicover, 1968): (*author*, *person who writes*), (*sicken*, *to make ill*), (*strong*, *potent (of tea)*). The decontextualization that reconstructs the verb in logical metonymy is just such an example of a dense paraphrase. We define a *dense paraphrase* as follows:

(6) **Definition 1: Dense Paraphrase**: Given the pair, (S,P), where S is a source expression, and P is an expression, we say P is a valid *dense paraphrase* of S if: P is an expression (lexeme, phrase, sentence) that eliminates any contextual ambiguity that may be present in S, but that also makes explicit, any underlying semantics that is not (usually) expressed in the economy of sentence structure, e.g., default or hidden arguments, dropped objects or adjuncts. P is both meaning preserving (consistent) and ampliative (informative) with respect to S.

The result of dense paraphrasing over an enriched compositional construction is, therefore, a surface textual realization of the covert semantic typing responsible for the apparent violation in selection. We see how this is realized computationally in the remainder of the paper.

Related Work

Language representation learning This task is to generate vector representations of natural language text that can be quantitatively analyzed (Naseem, Razzak, Khan, & Prasad, 2021). The word, as the basic unit of the text, has long been studied in representation learning. Early methods focus on categorical word representations, such as one-hot encoding, bag-of-words and TF-IDF (Jones, 2004), that can reflect the frequency and statistical distribution of word tokens from the text. To address the vector sparsity from categorical representations and generate text representations that are able to capture more syntactic and semantic information from the original text fragment, subsequent methods, including word2vec (Mikolov, Chen, Corrado, & Dean, 2013) and GloVe (Pennington, Socher, & Manning, 2014) use dense vectors of floating-point numbers to represent different words (word embeddings). These dense vectors are obtained from language models that are trained to determine the validity of a word sequence as a natural sentence.

While a word embedding is capable of capturing some aspects of the meaning of words, it cannot account for the context of each word, e.g., words with the same surface form are mapped to the same dense vector, even though they may have different interpretations under different contexts. To solve this, recent progress on large transformer-based language models such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), and T5 (Raffel et al., 2020) have been made to generate contextualized word embeddings where the same tokens have different meaningful representations from different contexts. This technique is useful to our work because of its ability to learn the word meanings dynamically from the entire context. We intend to explore whether it can also capture the nuances between predicates or arguments with similar surface forms from logical metonymy.

Logical metonymy Using computational methods to analyze the behavior of logical metonymy is still an underexplored area. Early research (Lapata & Lascarides, 2003; Shutova, 2009; Zarcone, Utt, & Padó, 2012) framed it as the problem to identify covert event candidates for the argument from the sentence. Probabilistic models and distributional semantic models have been applied to find event candidates that are semantically compatible with both the predicate verb and argument (McGregor, Jezek, Purver, & Wiggins, 2017). Given the simple sentence structures of the data that have been used[*], these methods mostly focused on finding high-typical event candidates that agree with the predicate, rather than identifying the specific covert event, because of the lack of broader context. A more recent work (Rambelli, Chersoni, Lenci, Blache, & Huang, 2020) compared the transformer-based and probabilistic models on their ability to interpret logical metonymy from the same data. Another work (Gietz & Beekhuizen, 2022) also applied the transformer-based model to perform a verb prediction task for complement coercion. However they focused on interpreting the coercion as a form of pragmatic enrichment. Compared with previous work, we explores the transformer model, namely BERT, in two modes: 1) we explore the semantic coercion capability of BERT on a newly curated set of sentences where logical metonymy exists. Our new data has more complex sentence structure and contains richer contexts; 2) we define a masked completion task for identifying the appropriate covert event and show the model can be improved on such tasks by exposing more explicitly expressed coercive sentences to it. More broadly, our task can be considered in the same vein with other semantic tasks that explore the implicit or underspecified components of a linguistic expression (Roth, Tsarfaty, & Goldberg, 2021).

Task Definition

We define our task as completing sentences where logical metonymy exists. The goal is to complete a sentence with its corresponding covert event while maintaining its syntactic correctness and semantic interpretability. We use the notion of masking for this completion task, where the `[MASK]` token is a proxy for dense paraphrasing. Table 1 shows the masked sentences with logical metonymy. The token `[MASK]` needs to be replaced by the covert event token in the correct form in a cloze style. This approach is an attempt to discover through the model and distributional behavior over the corpus what the likely type being selected by the predicate is.

Table 1: Examples of masked sentences. In each sentence, the **predicate** "coerces" its direct *noun phrase* into the event interpretation. Column EVENT lists the most plausible event of each sentence.

MASKED SENTENCE	EVENT
If you **enjoyed** `[MASK]` *this episode*, take a minute ...	listening to
The toughs calmly **finished** `[MASK]` *their beers*.	drinking
... that they would **stop** *bad things* `from [MASK]`.	happening

Data Preparation

In this section we describe the steps taken to build the dataset for the dense paraphrasing task: 1) we first collect the raw text from a large crawled textual corpus to ensure that the data we are using has a wide coverage of both text format and context that can be scraped from the web; 2) the data is further preprocessed by first filtering out irrelevant passages and then creating basic annotations from NLP pipelines; 3) finally we identify candidate sentences by checking if their semantic structures allow logical metonymy to exist.

For our task we select five verbs: the verb *enjoy* and four verbs from the aspectual class, i.e., verbs that denote a phase of an event and directly select for that event as their complement (*begin*, *continue*, *finish*, *stop*). These verbs are selected for two reasons: first, there is a vast literature that converges on the idea that these verbs may activate coercion in their object argument (Pustejovsky & Bouillon, 1995; Pulman, 1997); second, they are verbs that tend to be monosemous, which allows us to avoid performing verb sense disambiguation, that will be needed for other coercive classes such as perception verbs, *hear* and *listen*.

Data Collection

We draw our source materials from the Colossal Clean Crawled Corpus (C4) (Raffel et al., 2020), which is a collection of around 750 GB of English-language text based on the Common Crawl dataset. Web-crawled text data has been used

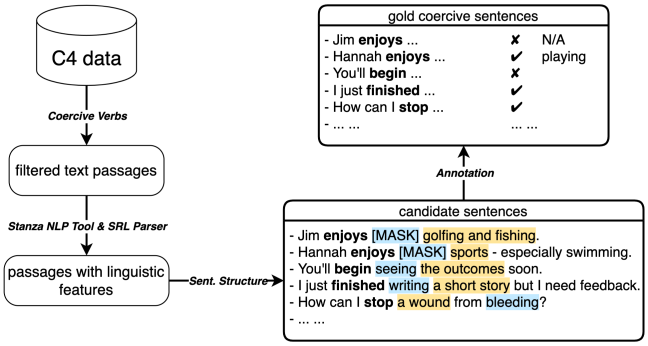
[*] Each sentence has the structure of *subject+verb+object*, e.g., *artist begins portrait*.

for pretraining in various NLP tasks including language modeling (Devlin et al., 2019), summarization (Rush, Chopra, & Weston, 2015), and machine translation (Luong, Pham, & Manning, 2015). In the latest, C4 is developed for training T5 (Raffel et al., 2020), a transformer-based model that can be applied to many text generation tasks. The extensive usage of C4 and the nature of it being sourced from the public Common Crawl web scrape ensure its huge volume and various language styles compared to other public datasets, which provides a "real world" text distribution for our task.

Preprocessing

We adopt a two-stage approach for the data preprocessing. First, given that only text with coercion are relevant to our topic, we extract the text pertaining to our task from the C4 by filtering out text spans that do not contain any coercive verbs we defined above. To make it efficient in computation, we take each passage (separated by the line break in each C4 document) as our basic text unit, and apply simple string match to check if a coercive verb or one of its inflections matches partial text from the passage. In this stage we can produce a coarse-grained subset of passages that match the coercive verbs. In the second stage, we run the Stanza pipeline (Qi, Zhang, Zhang, Bolton, & Manning, 2020) that includes sentence segmentation, tokenization, syntactic parsing, and dependency parsing over each selected passage to generate its basic linguistic features. Then we run the semantic role labeler (SRL) (Gardner et al., 2017) on each sentence to get the semantic roles.

In practice, considering the significant amount of bandwidth and computational power for downloading and preparing the dataset, we sample the C4 and apply the preprocessing approach on four batches of documents, which results in a set of 362,176 passages (3.5 GB) for further processing.



gold coercive sentences

- Jim **enjoys** ... ✘ N/A
- Hannah **enjoys** ... ✔ playing
- You'll **begin** ... ✘
- I just **finished** ... ✔
- How can I **stop** ... ✔
- ... ...

candidate sentences

- Jim **enjoys** [MASK] golfing and fishing.
- Hannah **enjoys** [MASK] sports - especially swimming.
- You'll **begin** seeing the outcomes soon.
- I just **finished** writing a short story but I need feedback.
- How can I **stop** a wound from bleeding?
- ... ...

Figure 1: Pipeline for extracting coercive sentences.

Coercion Sentence Extraction

While the preprocessing step can significantly reduce the amount of the data by filtering out text passages that do not contain any coercive verb, further steps are still required to extract real coercive sentences from the left passages. We adopt a semi-automatic approach for that purpose. Figure 1 shows the complete pipeline for extracting coercive sentences. First, C4 data is preprocessed into passages that contain coercive verbs with linguistic features. Then we examine the syntactic tags and semantic roles of each sentence to check if it complies to the coercive sentence structure we defined. Finally we use crowd-sourcing to check the plausibility of the coercive sentences and annotate the covert events.

Coercive Sentence Structure

We focus on the predicate-argument combination and define two sentence structures where coercion could happen as the templates to automatically extract data. In the predicate-argument pair where coercion exists, the desirable argument is always an NP. We utilize semantic roles and syntactic features including part-of-speech tags and universal dependencies to only retrieve the predicate-argument pairs where the arguments are NPs. For example, a sentence like *Adriana always enjoys when there is time for a good photoshoot.* is excluded because the direct argument of *enjoy* is a (subordinate) clause.

Coercion structure is for sentences where a coercion actually occurs. Each sentence in this structure has at least a semantic frame with a predicate-argument combination in which the predicate is a coercive verb and the argument is an NP (*coercive verb* + *argument*). Since the coercive verb directly "coerces" the argument, we require the argument to follow the verb immediately. Since sentences with this structure do not have a covert event expressed on the surface, we refer to them as the IMPLICIT dataset.

Decontextualized coercion structure Given the extensiveness of our source corpus, we also find a fair amount of data in which the supposedly hidden event verbs are explicitly shown in the surface structure which we refer to as the "decontextualization" of coercion. The structures of these sentences are defined as *coercive verb* + *event verb* + *argument* (or *coercive verb* + *argument* + *from* + *event verb* for ***stop***). We start the search by first finding predicate-argument pairs where the predicate is an event verb and the argument is still an NP. Later, we examine if there is a coercive verb that precedes the event verb to complete the triple. For the verb *stop*, we adopt a different structure where we first find the predicate-argument pair of an NP following a coercive verb immediately. Then we check if there is a "from *predicate*" pattern in which the predicate is an event verb. We refer to sentences with this type of structure as the EXPLICIT dataset, as the latent semantic information are expressed on the surface where the event verb is the mask token we want to reveal.

Table 2 shows the two strategies we design to introduce the notion of mask into these two datasets respectively. The first strategy is **insertion**: place the `[MASK]` token where the event verb is omitted, between the coercive verb and the argument. The second one is **replacing**: replace the event verb with the `[MASK]` token when it is explicit in the surface form.

Table 2: Examples of masking strategy

TYPE	RAW SENTENCE	MASKED SENTENCE
INSERT	Every morning, guests can enjoy a buffet breakfast.	Every morning, guests can enjoy `[MASK]` a buffet breakfast .
REPLACE	The basket was a nice touch and we enjoyed having the picnic style breakfast.	The basket was a nice touch and we enjoyed `[MASK]` the picnic style breakfast.

Annotating Coercive Sentences

Although previous steps can help find IMPLICIT and EXPLICIT sentences in the coercive structure. Human annotation is still required to address limitations of the datasets at this stage.

Presence of Coercion The IMPLICIT dataset requires manual judgment to decide whether the coercion is present indicating a mismatch between the type of the predicate and its argument, that is difficult to detect using only semantic and syntactic features. When an argument is a noun denoting a process, e.g., *golfing*, even though there is a predicate-argument combination with the coercive verb, it can be directly selected by the verb, hence no coercion. Therefore, if we forcefully insert a filler event verb between the coercive verb and the argument, the filler verb would not carry any additional semantic meaning.

Presence of Event Verb To train a model that can predict the hidden event verb in coercion, we need training data that has the event verb present. Unlike the EXPLICIT data which already has the event verb present, the IMPLICIT data still lacks the masked event that is not explicitly shown on the surface. The reason we are not satisfied with only using one type of dataset is that we hypothesize that the context provided by the two different types of data may vary in richness. Consider the following sentences:

(7) a. Alex finished **eating** the bread.
b. Alex finished the bread her husband baked.

Sentence (7a) has an explicit event verb *eating*, and the information of the event eating is conveyed directly by the word itself. If we mask or remove the event verb, the information associated with the "eating event" disappears, and there is no way for the model to learn from the context what the masked part could be, rather than guessing based on the co-occurrence of verbs and the argument *bread*. In contrast, the context in sentence (7b), which does not have an explicit event verb, tends to be more informative because the hidden information the event verb supposedly carries is transferred to other parts of the sentence. Our goal is to have the model learn the underlying semantics from the context, that the potential event is *eating*, as shown in (8), instead of *cooking* as the event of *baking* has appeared already and can be used as a semantic cue:

(8) Alex finished *[eating]* the bread her husband baked.

Annotation Schema We design a set of annotation schemas to leverage human judgments to tackle the above mentioned issues. We follow the insertion masking strategy and insert the `[MASK]` token either between the coercive verb and the argument or after the preposition *from* following the argument to reconstruct the sentence into its decontextualized form. The human annotators are asked to: 1) decide if coercion is present in the sentence; and 2) if coercion is present, provide a most plausible event verb given the context. Table 3 shows the statistics of our final dataset.

Table 3: # of sentences from each dataset per coercive verb.

	ENJOY	BEGIN	STOP	CONTINUE	FINISH	ALL
EXPLICIT	8,876	10,823	2,904	2,858	645	26,106
IMPLICIT	362	88	72	61	134	717

Experimental Design

We formulate the interpretation of semantic coercion or logical metonymy as a masked sentence completion task in which the mask indicates the latent semantic process in the coerced form. We use the `BERT-base-uncased` model as our baseline because of its ability of learning contextualized representations of the whole sentence. The pretraining paradigm for transformer-based models makes BERT a strong baseline for our task.

We also fine-tune the `BERT-base-uncased` model with our IMPLICIT and EXPLICIT coercive sentences to see if extra data can improve the model performance on our task. We first decontextualize the coercive sentences by keeping the covert event verb on the surface in the correct form. Then we fine-tune the BERT to ingest the semantic relations that underlie these sentences. This training objective aligns perfectly with the masked language modeling task for BERT pretraining, where random words in the input sentences are substituted by masks and the model will gradually learn to predict the masked token based on the surrounding context, so it is natural to frame our training task as a masked language modeling problem. We train BERT using our IMPLICIT dataset, EXPLICIT dataset, and one that combines both (COMBINED) where we randomly masked 20% of all the tokens. We discuss experiment results from the baseline and our fine-tuned models in the next section.

Table 4: Accuracy (%) of models tested on EXPLICIT

	BASELINE		FINE-TUNED-EXP.		
	Acc.@1	Acc.@3	Acc.@1	Acc.@3	Count
ENJOY	39.97	54.95	52.81	68.01	888
BEGIN	32.20	48.67	37.87	54.71	1093
CONTINUE	16.66	35.00	38.33	50.00	60
FINISH	21.52	35.41	33.33	47.56	288
STOP	28.72	43.97	31.91	46.80	282
ALL	32.93	48.52	41.82	57.48	2611

Table 5: Accuracy (%) of models tested on IMPLICIT

	Baseline		Fine-tuned-Imp.		Fine-tuned-Exp.		
	Acc.@1	Acc.@3	Acc.@1	Acc.@3	Acc.@1	Acc.@3	Count
ENJOY	32.83	58.20	70.14	92.53	47.76	77.61	67
BEGIN	31.25	56.25	62.50	68.75	50.00	68.75	16
CONTINUE	11.11	44.44	66.66	77.77	66.66	77.77	9
FINISH	35.00	35.00	55.00	70.00	45.00	70.00	20
STOP	44.44	50.00	27.77	61.11	33.33	55.55	18
ALL	33.07	52.30	60.76	80.76	46.92	72.30	130

Table 6: Accuracy (%) of models tested on COMBINED

	Baseline		Fine-tuned-Comb.		
	Acc.@1	Acc.@3	Acc.@1	Acc.@3	Count
ENJOY	39.47	55.18	52.77	69.21	955
BEGIN	32.19	48.78	38.05	55.27	1109
CONTINUE	15.94	36.23	42.02	53.62	69
FINISH	22.40	35.38	32.79	48.05	308
STOP	29.66	44.33	32.00	48.66	300
ALL	32.94	48.70	42.02	58.55	2741

Evaluation and Results

To assess the task and the utility of our data, we perform the evaluation on the models under different experiment settings. We use accuracy as it is the most common metric for evaluating masked language modeling (LM) tasks. Table 4 shows the performance of the baseline and the model fine-tuned on the EXPLICIT dataset. The Fine-Tuned-Exp. outperforms the Baseline on the overall and the individual verb performance for both accuracy@1 and accuracy@3. Table 5 shows the model results on the IMPLICIT testset. We compare the BASELINE with two models fine-tuned on the IMPLICIT (Fine-Tuned-Imp.) and EXPLICIT data, respectively. Both fine-tuned models outperform the Baseline, indicating the effectiveness of fine-tuning BERT for our task. For the verb *stop*, the accuracy@1 for Fine-Tuned-Imp. drops compared to Baseline. By examining the model output, we observed that a large proportion of the mismatch is due to predicting a semantically similar event verb instead of an exact match. For example, in sentence *He wrote about how a person's life is shaped by the world in which they live in, that there is nothing that one can do to* ***stop*** *events from [MASK], and that nature is indifferent.* The gold is ***occurring***; while the top two predictions ranked by probability generated by Fine-Tuned-Imp. model are, in order, ***happening*** and ***occurring***. If we simply focus on the top one candidate, the predictions are counted as a mismatch, but, in fact, the model understands the underlying semantic event and only ranks a synonym of the ground truth higher. When we look at the accuracy@3 on *stop*, the fine-tuned models attain higher numbers. It shows that increasing the candidate pool generated by the models would better reflect the model's ability to interpret coercion. Table 6 shows the model results on the COMBINED dataset. Similarly, Fine-Tuned-Combined outperforms the Baseline. Table 5 also shows that the Fine-Tuned-Imp. model achieves a higher accuracy than the Fine-Tuned-Exp. model on the IMPLICIT test set. This supports our hypothesis that the contexts in IMPLICIT and EXPLICIT differ in richness and may have an impact on the effectiveness of the training.

Table 7: Accuracy (%) of models tested on CE and L&L used in (Rambelli et al., 2020)

	Baseline		Fine-tuned-comb.		
	Acc.@1	Acc.@3	Acc.@1	Acc.@3	Count
CE	27.12	38.98	32.20	50.85	59
L&L	1.72	6.90	17.24	25.86	58
ALL	14.53	24.79	23.08	38.46	117

To check whether our fine-tuned model can generalize on new data, we apply the model on the data from (Rambelli et al., 2020). The two datasets they used, namely CE and L&L consist of a set of triples of subject, verb and object, and a set candidates events for a triple. We convert each triple to a masked sentence that can be consumed into our model, and use the candidate event with the highest probability as the gold answer. For example, the triple *(customer, start, dinner)* is converted to *customer starts [MASK] dinner* with the gold answer as *eating*. Table 7 shows the results on CE and L&L datasets. The low performance of the Baseline in our experiment further confirms their findings that interpreting logical metonymy is a challenging task. Compared to the BASELINE, however, the Fine-tuned-combined model improves the accuracy on both datasets by a large margin, indicating the transferability of our model to new coercive sentences. The overall result on CE and L&L is lower than that on our data (Table 6) due to the lack of context from the coercive triples, thus making it more difficult for the BERT model to generate the most plausible answers.

Discussion

In this paper, we show how the behavior of one class of type coercion, logical metonymy, can be captured with transformer-based architectures, using a theory of *dense paraphrasing*. Adopting the general view of compositionality outlined in (Pustejovsky, 1995), this theory proposes that sets of surface paraphrases act as the signature for a semantic type. When collected, these surface forms can be used as a dataset to then fine-tune an attention-based architecture.

For semantic theories that adopt "enriched compositional" mechanisms (Pustejovsky, 1995; Jackendoff, 1997), syntactic variation in argument position (polymorphism) is due to the application of covert (to the surface form) coercion operations that license syntactic realizations for a semantic type that is required by a predicate; i.e., the *canonical syntactic forms* for a semantic type.

Our results show that logical metonymy is a challenging task even to large pretrained models. However the results also show that the model can be substantially improved by fine-tuning through dense paraphrasing and generalize well on new data at the same time. We are currently expanding the model to include verb classes and coercion contexts well beyond the five pilot verbs studied here. The theoretical foundations for dense paraphrasing are developed in (Pustejovsky & Jezek, 2023).

References

Apresjan, J. D. (1974). Regular polysemy. *Linguistics*.

Asher, N. (2011). *Lexical meaning in context: A web of words*. Cambridge University Press.

Bhagat, R., & Hovy, E. (2013). What is a paraphrase? *Computational Linguistics*, *39*(3), 463–472.

Bu, F., Zhu, X., & Li, M. (2010). Measuring the non-compositionality of multiword expressions. In *Proceedings of the 23rd international conference on computational linguistics (coling 2010)* (pp. 116–124).

Choi, E., Palomaki, J., Lamm, M., Kwiatkowski, T., Das, D., & Collins, M. (2021). Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, *9*, 447–461.

Chomsky, N. (1965). *Aspects of the theory of syntax* (Vol. 11). MIT press.

Culicover, P. W. (1968). Paraphrase generation and information retrieval from stored text. *Mech. Transl. Comput. Linguistics*, *11*(3-4), 78–88.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Naacl-hlt.*

Dowty, D. (1979). *Word meaning and montague grammar: The semantics of verbs and times in generative semantics and in montague's ptq* (Vol. 7). Springer Science & Business Media.

Dowty, D. (1991). Thematic proto-roles and argument selection. *language*, *67*(3), 547–619.

Elazar, Y., Basmov, V., Goldberg, Y., & Tsarfaty, R. (2021). Text-based np enrichment. *arXiv e-prints*, arXiv–2109.

Fillmore, C. (2008). Frame semantics. In *Cognitive linguistics: Basic readings* (pp. 373–400). De Gruyter Mouton.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., … Zettlemoyer, L. S. (2017). Allennlp: A deep semantic natural language processing platform..

Gietz, F., & Beekhuizen, B. (2022, February). Remodelling complement coercion interpretation. In *Proceedings of the society for computation in linguistics 2022* (pp. 158–170). online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2022.scil-1.13`

Harris, Z. S. (1970). Transformational theory. In *Papers in structural and transformational linguistics* (pp. 533–577). Springer.

Holsinger, E. (2009). The effects of non-compositionality on language processing.

Jackendoff, R. (1972). *Semantic interpretation in generative grammar.* MIT Press.

Jackendoff, R. (1990). *Semantic structures* (Vol. 18). MIT press.

Jackendoff, R. (1997). *The architecture of the language faculty* (No. 28). MIT Press.

Jones, K. S. (2004). A statistical interpretation of term specificity and its application in retrieval. *J. Documentation*, *60*, 493-502.

Kahane, S. (1984). *The meaning-text theory.* De Gruyter.

Katz, J. J., & Fodor, J. A. (1963). The structure of a semantic theory. *language*, *39*(2), 170–210.

Lakoff, G. (1970). *Irregularity in syntax*. Holt, Rinehart, and Winston.

Lapata, M., & Lascarides, A. (2003). A probabilistic account of logical metonymy. *Computational Linguistics*, *29*(2), 261–315. Retrieved from `https://aclanthology.org/J03-2004` doi: 10.1162/089120103322145324

Luong, T., Pham, H., & Manning, C. D. (2015, September). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1412–1421). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D15-1166` doi: 10.18653/v1/D15-1166

McGregor, S., Jezek, E., Purver, M., & Wiggins, G. (2017). A geometric method for detecting semantic coercion. In *Iwcs 2017-12th international conference on computational semantics-long papers.*

Melamed, I. D. (1997). Automatic discovery of non-compositional compounds in parallel data. *arXiv preprint cmp-lg/9706027*.

Mel'čuk, I. (1995). Phrasemes in language and phraseology in linguistics. *Idioms: Structural and psychological perspectives*, 167–232.

Mel'čuk, I. (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of phraseology*, *3*(1), 31–56.

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Iclr.*

Naseem, U., Razzak, I., Khan, S. K., & Prasad, M. (2021). A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, *20*, 1 - 35.

Partee, B. (1973). Some transformational extensions of montague grammar. *Journal of Philosophical Logic*, *2*(4), 509–534.

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Doha, Qatar: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D14-1162` doi: 10.3115/v1/D14-1162

Pollard, C., & Sag, I. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Pulman, S. G. (1997). Aspectual shift as type coercion. *Transactions of the Philological Society*, *95*(2), 279–317.

Pustejovsky, J. (1995). *The generative lexicon*. Cambridge, MA: MIT Press.

Pustejovsky, J. (2012). Co-compositionality in grammar. *The Oxford handbook of compositionality*, *371*, 382.

Pustejovsky, J., & Bouillon, P. (1995). Aspectual coercion and logical polysemy. *Journal of semantics*, *12*(2), 133–162.

Pustejovsky, J., & Jezek, E. (2008). Semantic coercion in language: Beyond distributional analysis. *Italian Journal of Linguistics*, *20*(1), 175–208.

Pustejovsky, J., & Jezek, E. (2023). *Generative Lexicon Theory: A Modern Introduction.* Oxford University Press.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. In *Acl.*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1-67. Retrieved from `http://jmlr.org/papers/v21/20-074.html`

Rambelli, G., Chersoni, E., Lenci, A., Blache, P., & Huang, C.-R. (2020, December). Comparing probabilistic, distributional and transformer-based models on logical metonymy interpretation. In *Proceedings of the 1st conference of the asia-pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing.*

Roth, M., Tsarfaty, R., & Goldberg, Y. (Eds.). (2021, August). *Proceedings of the 1st workshop on understanding implicit and underspecified language.* Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.unimplicit-1.0`

Rush, A. M., Chopra, S., & Weston, J. (2015, September). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 379–389). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D15-1044` doi: 10.18653/v1/D15-1044

Self. (n.d.). Dense paraphrasing for textual enrichment. *Anon.*

Shutova, E. (2009, August). Sense-based interpretation of logical metonymy using a statistical method. In *Proceedings of the ACL-IJCNLP 2009 student research workshop* (pp. 1–9). Suntec, Singapore: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/P09-3001`

Smaby, R. (1971). *Paraphrase grammars, volume 2 of formal linguistics series. dordrecht: D.* Reidel Publishing Company.

Van Valin, R. D. (1999). Generalized semantic roles and the syntax-semantics interface. *Empirical issues in formal syntax and semantics*, *2*, 373–389.

Wu, Z., Luan, Y., Rashkin, H., Reitter, D., & Tomar, G. S. (2021). Conqrr: Conversational query rewriting for retrieval with reinforcement learning. *arXiv preprint arXiv:2112.08558.*

Yazdani, M., Farahmand, M., & Henderson, J. (2015, September). Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1733–1742). Lisbon, Portugal: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D15-1201` doi: 10.18653/v1/D15-1201

Zarcone, A., Utt, J., & Padó, S. (2012, June). 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)* (pp. 70–79). Montréal, Canada: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/W12-1707`