IMPLI: Investigating NLI Models' Performance on Figurative Language

Kevin Stowe[◊] **Prasetya Ajie Utama**[†][*] **Iryna Gurevych**[◊]

[◊] Ubiquitous Knowledge Processing Lab (UKP)
Department of Computer Science
Technical University of Darmstadt
[†] Bloomberg, London, United Kingdom
www.ukp.tu-darmstadt.de

Abstract

Natural language inference (NLI) has been widely used as a task to train and evaluate models for language understanding. However, the ability of NLI models to perform inferences requiring understanding of figurative language such as idioms and metaphors remains understudied. We introduce the IMPLI (Idiomatic and Metaphoric Paired Language Inference) dataset, an English dataset consisting of paired sentences spanning idioms and metaphors. We develop novel methods to generate 24k semi-automatic pairs as well as manually creating 1.8k gold pairs. We use IMPLI to evaluate NLI models based on RoBERTa fine-tuned on the widely used MNLI dataset. We then show that while they can reliably detect entailment relationship between figurative phrases with their literal counterparts, they perform poorly on similarly structured examples where pairs are designed to be non-entailing. This suggests the limits of current NLI models with regard to understanding figurative language and this dataset serves as a benchmark for future improvements in this direction.[1]

1 Introduction

Understanding figurative language (i.e., that in which the intended meaning of the utterance differs from the literal compositional meaning) is a particularly difficult area in NLP (Shutova, 2011; Veale et al., 2016), but is essential for proper natural language understanding. We consider here two types of figurative language: idioms and metaphors. Idioms can be viewed as non-compositional multiword expressions (Jochim et al., 2018), and have been historically difficult for NLP systems. For instance, sentiment systems struggle with multiword expressions in which individual words do not directly contribute to the sentiment (Sag et al., 2002).

Idioms	Jamie was *pissed off* this afternoon. → Jamie was *irritated* this afternoon
	There's a marina down *in the docks*. ↛ There's a marina down *under scrutiny*.
Metaphors	The *hearts of men were softened*. → The *men were made kindler and gentler*.
	The gun *kicked* into my shoulder. ↛ The *mule* kicked into my shoulder.

Table 1: Examples of entailment (→) and non-entailment pairs (↛) from the IMPLI dataset.

Metaphors involve linking conceptual properties of two or more domains, and are known to be pervasive in everyday language (Lakoff and Johnson, 1980; Stefanowitsch and Gries, 2008; Steen et al., 2010). Recent work has shown that these types of figurative language are impactful across a broad array of NLP tasks (see §2.1).

Large-scale pre-training and transformer-based architectures have yielded increasingly powerful language models (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019). However, relatively little work has explored these models' representations of figurative and creative language. NLI datasets have widely been used for evaluating the performance of language models (Dagan et al., 2006; Bowman et al., 2015a; Williams et al., 2018), but there are insufficient figurative language datasets in which a literal sentence is linked to a corresponding figurative counterpart that are large enough to be suitable for evaluating NLI. Due to the creative nature of human language, creating a dataset of diverse, high-quality literal/figurative pairs is time-consuming and difficult.

To address this gap, we build a new English dataset of paired expressions designed to be leveraged to explore model performance via NLI. Our dataset, IMPLI (Idiomatic/Metaphoric Paired Language Inference), is comprised of both silver pairs, which are built using semi-automated

[*] The work was done while the second author was still affiliated with the UKP Lab at TU Darmstadt.

[1] Dataset and all related resources are publicly available at https://github.com/UKPLab/acl2022-impli.

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5375 - 5388
May 22-27, 2022 ©2022 Association for Computational Linguistics

methods (§3.1), as well as hand-written gold pairs (§3.4), crafted to reflect both entailment and non-entailment scenarios. Each pair consists of a sentence containing a figurative expression (idioms/metaphors) and a literal counterpart, designed to be either entailed or non-entailed by the figurative expression (Table 1 shows some examples).

Our contribution thus consists of three key parts:

- We create a new **IMPLI** dataset consisting of 24,029 silver and 1,831 gold sentence pairs consisting of idiomatic and metaphoric phrases that result in both entailment and non-entailment relationship (see Table 2).
- We evaluate language models in an NLI setup, showing that metaphoric language is surprisingly easy, while non-entailing idiomatic relationships remain extremely difficult.
- We evaluate model performance in a number of experiments, showing that incorporating idiomatic expressions into the training data is less helpful than expected, and that idioms that can occur more in more flexible syntactic contexts tend to be easier to classify.

2 Background

2.1 Figurative Language and NLP

Figurative language includes idioms, metaphors, metonymy, hyperbole, and more. Critically, figurative language is that in which speaker meaning (what the speaker intends to accomplish through an utterance) differs from the literal meaning of that utterance. This leads to problems in NLP systems if they are trained mostly on literal data, as their representations for particular words and/or phrases will not reflect their figurative intended meanings.

Figurative language has a significant impact on many NLP tasks. Metaphoric understanding has been shown to be necessary for proper machine translation (Mao et al., 2018; Mohammad et al., 2016). Sentiment analysis also relies critically on figurative language: irony and sarcasm can reverse the polarity of a sentence, while metaphors and idioms may make more subtle changes in the speaker meaning (Ghosh et al., 2015). Political discourse tasks including bias, misinformation, and political framing detection benefit from joint learning with metaphoricity (Huguet Cabot et al., 2020). Figurative language engendered by creativity on social media also poses difficulty for many NLP tasks including identifying depression symptoms (Yadav et al., 2020; Iyer et al., 2019) and hate speech detection (Lemmens et al., 2021).

We are here focused on idioms and metaphors. There is currently a gap in diagnostic datasets for idioms, and our work fills this gap. There exist some relevant metaphoric resources (see §2.2); metaphors are known to be extremely common and important to understanding figurative language, our resource serves to build upon this work.

2.2 NLI and related challenges

Natural language inference is the task of predicting, given two fragments of text, whether the meaning of one (*premise*) entails the other (*hypothesis*) (Dagan et al., 2006). The task is formulated as a 3-way classification problem, in which the premise and hypothesis pairs are labeled as *entailment*, *contradiction*, or *neutral*, if their relationship could not be directly inferred (Bowman et al., 2015b). NLI has been widely used as an evaluation task for language understanding, and there have been a large number of challenging datasets, which have been used to further our understanding of the capabilities of language models (Wang et al., 2018, 2019).

Paired data for figurative language is relatively sparse, and there is a gap in the diagnostic datasets used for NLI in this area. Previous work includes the literal/metaphoric paraphrases of Mohammad et al. (2016) and Bizzoni and Lappin (2018), although both contain only hundreds of samples, insufficient for proper model training and evaluation. With regard to NLI, early work proposed the task of textual entailment as a way of understanding metaphor processing capabilities (Agerri et al., 2008; Agerri, 2008). Poliak et al. (2018) build a dataset for diverse NLI, which includes some creative language such as puns, albeit making no claims with regard to figurativeness.

Zhou et al. (2021) build a dataset consisting of paired idiomatic and literal expressions. They begin with a set of 823 idiomatic expressions yielding 5,170 sentences, and had annotators manually rewrite sentences containing these idioms as literal expressions. We expand on this methodology by having annotators only correct definitions for the idioms themselves and use these definitions to automatically generate the literal interpretations of the idioms by replacing them into appropriate contexts: this allows us to scale up to over 24k silver sentences. We also expand beyond paraphrasing by incorporating both entailment and non-entailment

Fig. Type	Ent.	Gold/silver	Description	Count
Idioms	→	Silver	Replace idiom used in figurative context with definition	16652
	↛	Silver	Replace idiom used in literal context with definition	886
	↛	Silver	Replace idiom used in figurative context with adversarial definition	6116
	→	Gold	Hand written literal definition of idiom	532
	↛	Gold	Manual replacement of key words in definition w/ antonyms	375
	↛	Gold	Hand written non-entailed sentence	254
Metaphors	→	Silver	Replace metaphoric construction with literal construction	375
	→	Gold	Hand written literal paraphrase of metaphor	388
	↛	Gold	Hand written non-entailed sentence	282

Table 2: **Dataset Summary**: Overview of entailments/non-entailment types in IMPLI. (→) denotes entailments, (↛) non-entailments. Note that the descriptions are simplified: some intermediate steps are omitted (see §3.1).

pairs to enable NLI-based evaluation.

Similar to this work, Chakrabarty et al. (2021a) build a dataset for NLI based on figurative language. Their dataset consists of figurative/literal pairs recast from previously developed simile and metaphor datasets, along with a parallel dataset between ironic and non-ironic rephrasing. This sets the groundwork for figurative NLI, but the dataset is relatively small outside of the irony domain, and the non-entailments are generated purely by replacing words with their antonyms, restricting the novelty of the hypotheses. Their dataset is relatively easy for NLI models; here we show that figurative language can be challenging, particularly with regard to non-entailments.

Zhou et al. (2021) and Chakrabarty et al. (2021a) provide invaluable resources for figurative NLI; our works aims to covers gaps in a number of areas. First, we generate a large number of both entailment and non-entailment pairs, allowing for better evaluation of adversarial non-entailing examples. Second, our silver methods allow for rapid development of larger scale data, allowing for model training and evaluation. We show that while entailment pairs are relatively easy (accuracy scores ranging from .86 to .89), the non-entailment pairs are exceedingly challenging, with the `roberta-large` model achieving accuracy scores ranging from .311 to .539.

3 Building a Dataset

Our IMPLI dataset is built from idiomatic and metaphoric sentences paired with entailing and non-entailing counterparts, from both silver pairs (§3.1) and manually written sentences (§3.4). For our purposes, we follow McCoy et al. (2019) in conflating the neutral and contradiction categories into a non-entailment label. We then label every pair as either entailment (→) or non-entailment (↛).

Due to the difficult nature of the task and to avoid issues with crowdsourcing (Bowman et al., 2020), we employed expert annotators. We used two fluent English speakers, both graduate students in linguistics with strong knowledge in figurative language, paid at a rate of $20/hr. For each method below, we ran pilot studies, incorporated annotator feedback and iteratively assessed the viability of identifying and generating appropriate expressions. As the annotators were working on generating new expressions, agreement was not calculated: we instead assessed the quality of the resulting expressions (see Section 3.3). Table 2 contains an overview of the different entailment and non-entailment types collected (Detail examples are also provided in Appendix D).

3.1 Silver pairs

First, we explore a method for generating silver pairs using annotators to create phrase definitions which can be inserted automatically into relevant contexts, yielding a large number of possible entailment and non-entailment pairs that differ only with regard to the relevant phrase. Our procedure hinges on a key assumption: for any given figurative phrase, we can generate a contextually independent literal paraphrase. We then replace the original expression with the literal paraphrase, following the assumption that the figurative expression necessarily entails its literal paraphrase:

> He's stuck in bed, which is his *hard cheese*. → He's stuck in bed, which is his *bad luck*.

Conversely, in contexts where the original phrase is used literally, replacing it with the literal paraphrase should yield a non-entailment relation.

> Switzerland is famous for six cheeses, sometimes referred to as *hard cheeses*.

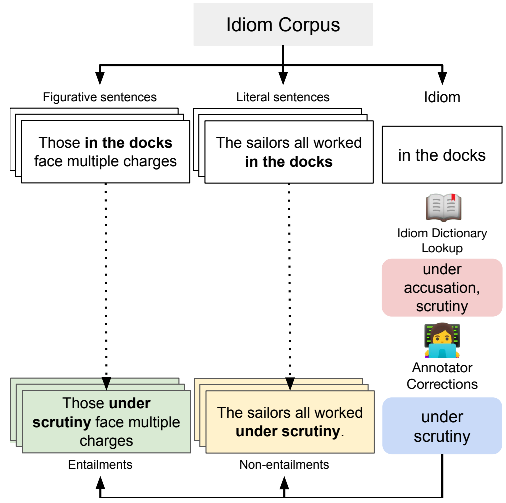
Idiom Corpus
Figurative sentences
Literal sentences
Idiom
Those in the docks face multiple charges
The sailors all worked in the docks
in the docks
Idiom Dictionary Lookup
under accusation, scrutiny
Annotator Corrections
under scrutiny
Those under scrutiny face multiple charges
The sailors all worked under scrutiny.
Entailments
Non-entailments

Figure 1: **Idiomatic definition replacement**. Pairs are generated using corrected dictionary definitions, substituted into figurative (left) and literal (center) sentences.

↛ Switzerland is famous for six cheeses, sometimes referred to as *bad luck*.

3.1.1 Idioms

To build idiomatic pairs, we use three corpora that contain sentences with idiomatic expressions (IEs) labelled as either figurative or literal.[2] These are the MAGPIE Corpus (Haagsma et al., 2020), the PIE Corpus (Adewumi et al., 2021), and the SemEval 2013 Task 5 (Korkontzelos et al., 2013). We collect the total set of IEs that are present in these corpora. We then extract definitions for these using freely available online idiom dictionaries.[3]

These definitions are often faulty, incomplete, or improperly formatted. We employed annotators to make manual corrections. The annotators were given the original IE as well as the definition extracted from the dictionary. The annotators were asked to ensure that the dictionary definition given was (1) a correct literal interpretation and (2) fit syntactically in the same environments as the original IE. If the definition met both of these criteria, the IE can be replaced by its definition to yield an entailment pair. If either criterion was not met, annotators were asked to minimally update the definition so that it satisfied the requirements.

In total this process yielded 697 IE definitions. We then used the above corpora, replacing these definitions into the original sentences (see Figure 1). We use the figurative/literal labels from the original corpora: replacing them into figurative contexts yields entailment relations, while replacing them into contexts where the phrase is meant literally then yields non-entailments.

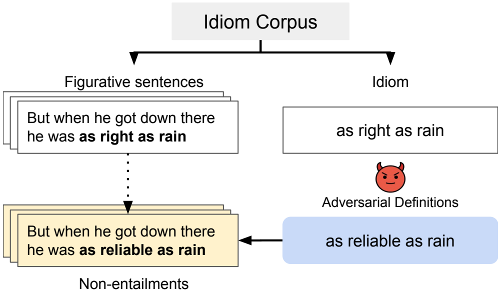

Figure 2: **Adversarial Pair Generation**. Non-entailing pairs are generated by replacing adversarial definitions into figurative contexts.

Original IE	Adversarial Definition
man of the cloth	tailor
heart of gold	cold, mean heart
come clean	bathe
turn a trick	do a magic trick

Table 3: Sampled hand-written adversarial definitions.

3.1.2 Adversarial Definitions

As a second method for generating non-entailment pairs, we asked annotators to write novel, adversarial definitions for IEs. Given a particular phrase, they were instructed to invent a new meaning for the IE that was not entailed by the true meaning, but which seemed reasonable presuming they had never heard the original IE. Some examples of this process are shown in Table 3.

We then replace these adversarial definitions into figurative sentences from the corpora. This yields pairs where the premise is an idiom used figuratively, and the hypothesis is a sentence that attempts to rephrase the idiom literally, but does so incorrectly, thus yielding non-entailments (Figure 2).

3.1.3 Metaphors

Metaphors are handled in a similar way: we start with a collection of minimal metaphoric expressions (MEs). These are subject-verb-object and adjective-noun constructions from Tsvetkov et al. (2014). Each is annotated as being either literal or metaphoric, along with an example sentence. We passed these MEs directly to annotators, who were then instructed to replace a word in the ME so that it would be considered literal in a neutral context.

1. *drop* prices → *reduce* prices

[2] We here use "idiomatic expression" or "IE" to refer to the specific idiom in question (ie. "kick the bucket", "spill the beans"), as opposed to the sentence/context containing it.

[3] www.theidioms.com, www.wiktionary.org

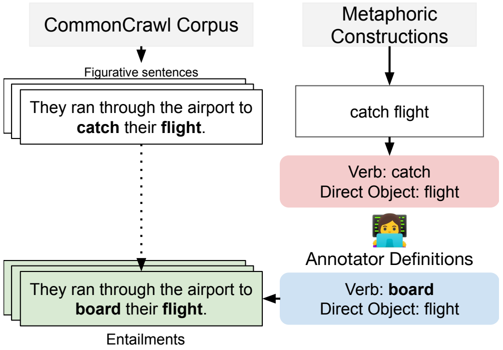

Figure 3: **Metaphor entailment generation**. Pairs are generated using annotator-defined literal translations substituted into metaphoric contexts.

2. **hard** truth → **unpleasant** truth
3. *hairy* problem → *difficult* problem

These can then be replaced in a similar fashion: we start with the original figurative sentence, replace the ME with the literal replacements, and the result is an entailing pair with the metaphoric sentence entailing the literal.

We apply this procedure to the dataset of Tsvetkov et al. (2014), yielding 100 metaphoric/literal NLI entailment pairs. We then take a portion of the Common Crawl dataset[4], and identify sentences that contain these original MEs. We identify sentences that contain the words from the metaphoric phrase, and replace the metaphoric word itself with its literal counterpart. This yields 645 additional silver pairs.

3.2 Postprocessing

For all silver methods, we also employ syntactic postprocessing to overcome a number of hurdles. First, phrases used idiomatically often follow different syntactic patterns than when used literally.

> **Original:** These point *out of this world*, but where to is not made clear.
> **Replaced:** *These point *wonderful*, but where to is not made clear.

This phrase in literal contexts functions syntactically as a prepositional phrase, while idiomatically it is used as an adjective. When replaced with the definition "wonderful" in a literal context, we get a grammatically incoherent sentence. Second, phrases in their literal usage often do not form full constituents, due to the string-matching approach of the original datasets. Many literal usages of these phrases are thus incompatible with the defined replacement.

- I think [this one has *to die*] *for* the other one to live.
- Turn *in* [*the raw* edges] of both seam allowances towards each other and match the folded edges.

To avoid these issues, we ran syntactic parsing on the definition and the expression within each context, requiring that the expression in context begins with the same part of speech as the definition and that it does not end inside of another phrase.

Additionally, for each replacement, we ensured that the verb conjugation matched the context. For this, we identified the conjugation in the context, and used a de-lemmatization script to conjugate the replacement verb to match the original.

3.2.1 Additional Issues

In implementing and analyzing this procedure, we noted a number of practical issues. First, a large number of the MEs provided are actually idiomatic or proverbial: the focus word does not actually contribute to the metaphor, but rather the entire expression is necessary. Similarly, we found that replacing individual parts of MEs is often insufficient to fully remove the metaphoric meaning. We iterated over possible solutions to circumvent these issues and found that it is best to simply skip instances for which a replacement does not yield a feasible literal interpretation.

3.3 Evaluating Pair Quality

In order for these automatically created pairs to be useful for NLI-based evaluation, they need to be of sufficiently high quality. As the annotators were generating novel definitions and pairs, rather than inter-annotator agreement, we instead evaluate the quality of the resulting pairs by testing whether the automatically generated pairs contained the appropriate entailment relation. For this task, each annotator was given 100 samples for each general category of silver generations (idiomatic entailments, idiomatic non-entailments, and metaphoric entailments). They were asked if the entailment relation between the two sentences was as expected. An expert than adjudicated disagreements to determine the final percentage of valid pairs.

To evaluate the syntactic validity of the generated pairs, we additionally ran the Stanford PCFG dependency parser (Klein and Manning, 2003) on

[4] https://commoncrawl.org/

	→ Idioms	↛ Idioms	→ Met.
Correct Entailments	%88	%90	%97
Premise S root	%89	%90	%82
Hypothesis S root	%90	%90	%82

Table 4: **Valid pairs**. Percentage of valid pairs, syntactically and with regard to the intended entailments, of automatic data generation.

the pairs. Per previous work in NLI (Williams et al., 2018), we evaluate the proportion of sentences for which the root node is S.

Table 4 shows the results. The semi-supervised examples evoked the correct entailment relation between %88 and %97 of the time: while there is still noise present, this indicates the effectiveness of the proposed methods. With regard to syntax, we see S node roots for between 82% and %90 of the sentences: within the range of the SNLI performance (74%-88%), and slightly behind the MNLI (91%-98%). We find that the generated hypotheses are not significantly different in quality than the premises. This indicates that the method for generation preserves the original syntax.

These methods allow us to quickly generate a substantial number of high-quality pairs to evaluate NLI systems on figurative language. However, they may introduce additional bias as we employ a number of restrictions in order to ensure syntactic and semantic compatibility, and we lack full non-entailment pairs for metaphoric data. We therefore expand our dataset with manually generated pairs.

3.4 Manual Creation of Gold Pairs

To create gold pairs, annotators were given a figurative sentence along with the focus of the figurative expression: for idioms, this is the IE; for metaphors, the focus word of the metaphor. For idioms, we used the MAGPIE dataset to collect contextually figurative expressions. For metaphors, we collected metaphoric sentences from the VUA Metaphor Corpus (Steen et al., 2010), the metaphor dataset of (Mohammad et al., 2016), and instances from the Gutenberg poetry corpus (Jacobs, 2018) annotated for metaphoricity (Chakrabarty et al., 2021b; Stowe et al., 2021) . Annotators were instructed to rewrite the sentence literally. This was done by removing or rephrasing the figurative component of the sentence. This yields gold standard paraphrases for idiomatic and metaphoric contexts.

We then asked annotators to write non-entailed hypotheses for each premise. They were encouraged to keep as much of the original utterance as possible, ensuring high lexical overlap, while removing the main figurative element of the sentence. For idioms, this comes from adding or adjusting words to force a literal reading of the idiom:

- The old girl finally kicked the bucket. ↛ The girl kicked the bucket on the right.

For metaphors, this typically involves keeping the same phrasing while adapting the sentence to have a different, non-metaphoric meaning.

- You must adhere to the rules. ↛ You must adhere the rules to the wall.

3.5 Antonyms

Previous work in NLI has employed the technique of replacing words in the literal sentences with their antonyms to yield non-entailing pairs (Chakrabarty et al., 2021a). We replicate this process for idioms: for the manually elicited definitions, we replace key words as determined by annotators with their antonyms. This yields sentences which negate the original figurative meaning and are thus suitable non-entailment pairs. Previous work found this antonym replacement for figurative language remains relatively easy for NLI systems, which we can additionally explore with regard to idioms.

These manual annotations provide a number of concrete benefits. First, they are not restricted to individual words or phrases (excluding antonyms): the figurative components can be rewritten freely, allowing for diverse, interesting pairs. Second, they are written by experts, ensuring higher quality than the automatic annotations, which may be noisy.

4 Experiments / Results

Using the `IMPLI` dataset, we aim to answer a series of questions via NLI pertaining to language models' ability to understand and represent figurative language accurately. These questions are:

- R1: *How well do pre-trained models perform on figurative entailments and non-entailments?*
- R2: *Does adding idiomatic pairs into the training data affect model performance?*
- R3: *Does the flexibility of idiomatic expressions affect model performance?*

Our dataset provides unique advantages in addressing these research questions that cover gaps

			Idioms						Metaphors		
Model	MNLI	MNLI-MM	$\rightarrow$ S	$\nrightarrow$ S^l	$\nrightarrow$ S^d	$\rightarrow$ G	$\nrightarrow$ G^a	$\nrightarrow$ G	$\rightarrow$ S	$\rightarrow$ G	$\nrightarrow$ G
`roberta-base`	.878	.876	.848	.539	.409	.890	.771	.311	.947	.818	.818
`roberta-large`	.899	.899	.866	.536	.418	.889	.777	.348	.936	.871	.840

Table 5: **R1: Model accuracy**. Accuracy on MNLI and **`IMPLI`** pairs, divided into silver (S) and gold (G) datasets. S^l Silver non-entailment based on replacement in literal contexts, S^d Silver non-entailment based on adversarial definitions, G^a Gold non-entailment based on antonyms.

in previous work: it contains a large number of both entailments and non-entailments and is large enough to be used for training the models.

R1: pre-trained Model Performance

We obtain baseline NLI models by fine-tuning `roberta-base` and `roberta-large` models on the MNLI dataset (Williams et al., 2018), with entailments as the positive class and all others as the negative and evaluate them on their original test sets as well as **`IMPLI`**.[5] Due to variance in neural model performance (Reimers and Gurevych, 2017), we take the mean score over 5 runs using different seeds.

We report results in Table 5. We observe that idiomatic entailments are relatively easy to classify, with accuracy scores over .84. Non-entailments were much more challenging. Silver pairs generated through adversarial definitions were especially difficult: the pairs contain high lexical overlap, and in many cases the premise and hypotheses are semantically similar. The replacement into literal samples were easier, as the idiomatic definition clashes more starkly with the original premise, making non-entailment predictions more likely. Consistent with Chakrabarty et al. (2021a)'s work in metaphors, non-entailment through antonym replacement is easiest for idioms: the antonymic relationship can be a marker for non-entailment, despite the high word overlap.

With regard to metaphors, silver entailment pairs are relatively easy. Manual pairs are more challenging but are still much easier than idioms. This is supported by the fact that metaphors are common in everyday language: these models have likely seen the same (or similar) metaphors in training. Our findings show that in fact metaphoricity may not be particularly challenging for deep pre-trained models, as they are able to effectively capture the metaphoric entailment relations. The `roberta-large` model performs better for metaphoric expressions than `roberta-base`, but the difference on other partitions is relatively small.[6] We also find that lexical overlap plays a significant role here as noted by previous work (McCoy et al., 2019): sentences with high overlap tend to be classified as entailments regardless of the true label (for more, see Appendix B).

We note that the manual pairs tend to be more difficult for both idioms and metaphors: these pairs can be more flexible and creative, whereas the silver pairs are restricted to more regular patterns.

R2: Incorporating Idioms into Training To evaluate incorporating idioms into training, we then split the idiom data by idiomatic phrase types, keeping a set of IEs separate as test data to assess whether the model can learn to correctly handle novel, unseen phrases. Our goal is to assess whether poor performance is due to models' not containing these expressions in training, or because their ability to represent figurative language inherently limited. We hypothesize that the non-compositional nature of these types of figuration should lead to poor performance on unseen phrases, even if the model is trained on other idiomatic data.

For each task, we split the data into 10 folds by IE and incrementally incorporate these folds into the original MNLI for training, leaving one fold out for testing. We experiment with incorporating all training data for both labels, as well as using only entailment or non-entailment samples. We then evaluate our results on the entire test set, as well as the entailment and non-entailment partitions.

Figure 4 shows the results, highlighting that additional training data yields only small improvements. Pairs with non-entailment relations remain exceedingly difficult, with performance capping out at only slightly better than chance. As hypothesized, additional training data is only somewhat effective in improving language models' idiomatic capabilities; this is not sufficient to overcome difficulties from literal usages of idiomatic phrases and adversarial definitions, indicating that idiomatic

[5] Model hyperparameters found in Appendix A.

[6] We found minimal differences between these models across R1-R3.

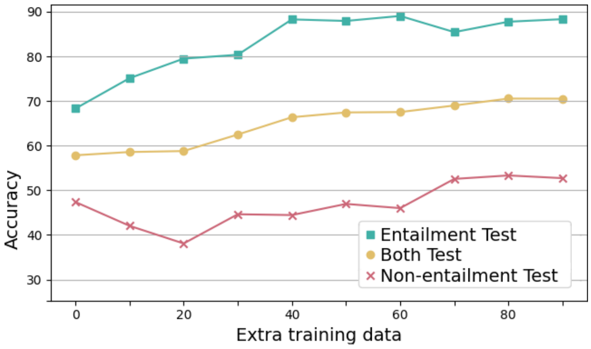

Figure 4: **R2: Training**. Performance of the `roberta-base` models as more idiom examples are added to the training data.

language remains difficult for pre-trained language models to learn to represent.

R3: Syntactic Flexibility Finally, we assess models' representation of idiomatic compositionality. Nunberg et al. (1994) indicate that there are two general types of idioms: "idiomatic phrases", which exhibit limited flexibility and generally occur only in a single surface form, and "idiomatically combining expressions" or ICEs, in which the constituent elements of the idiom carry semantic meaning which can influence their syntactic properties, allowing them to be more syntactically flexible.

For example, in the idiom *spill the beans*, we can map the spilling activity to divulging of information, and the beans to the information. Because this expression has semantic mappings to figurative meaning for its syntactic constituents, Nunberg et al. (1994) argue that it can be more syntactically flexible, allowing for expressions like *the beans that were spilled by Martha* to maintain idiomatic meaning. For fixed expressions such as *kick the bucket*, no syntactic constituents map directly to the figurative meaning ("die"). We then expect less syntactic flexibility, and thus *the bucket that was kicked by John* loses its idiomatic meaning.

We hypothesize that model performance will be correlated with the degree to which a given idiom type is flexible: more fixed expressions may be easier, as they are seen in regular, fixed patterns that the models can memorize, while more flexible ICEs will be more difficult, as they can appear in different patterns, cases, and word order, often even mixing in with other constituents. To test this, we define an ICE score as the percentage of times a phrase occurs in our test data in a form that does not match its original base form. Higher percentages mean the phrase occurs more frequently in a non-standard form, acting as a measure for the syntactic flexibility of the expression. We assessed the performance of the `roberta-base` model for each idiom type, evaluating Spearman correlations between performance and idioms' ICE scores.

We found no correlation between ICE scores and performance for entailments, nor for adversarial definition non-entailments (r = .004/.45, p = .921/.399, see Appendix C). However, we do see a weak but significant correlation (r = .188, p = 0.016) with non-entailments from literal contexts: the model performs better when the phrases are more flexible, contrary to our initial hypothesis.

One possible explanation is that the model memorizes a specific figurative meanings for each fixed expression, disregarding the possibility of these words being used literally. When the expression is used in a literal context, the model then still assumes the figurative meaning, resulting in errors on non-entailment samples. The ICEs are more fluid, and thus the model is less likely to have a concrete representation for the given phrase: it is better able to reason about the context and interacting words within the expression, making it easier to distinguish the entailing and non-entailing samples.

5 Conclusions and Future Work

In this work, we introduce the **`IMPLI`** dataset, which we then use to evaluate NLI models' capabilities on figurative language. We show that while widely used MNLI models handle entailment admirably and metaphoric expressions are relatively easy, non-entailment idiomatic relationships are more difficult. Additionally, adding idiom-specific training data fails to alleviate poor performance for non-entailing pairs. This highlights how currently language models are inherently limited in representing some figurative phenomena and can provide a target for future model improvements.

For future work, we aim to expand our data collection processes to new data sources. Our dataset creation procedure relies on annotated samples and definitions: as more idiomatic and metaphoric resources become available, this process is broadly extendable to create new figurative/literal pairs. Additionally, we only explore this data for evaluating NLI systems: this data could also be used for other parallel data tasks such as figurative language interpretation (Shutova, 2013; Su et al., 2017) and figurative paraphrase generation. As natural language generation often relies on training or fine-tuning models with paired sentences, this data could be a valuable resource for figurative language generation systems.

References

Tosin P. Adewumi, Saleha Javed, Roshanak Vadoodi, Aparajita Tripathy, Konstantina Nikolaidou, Foteini Liwicki, and Marcus Liwicki. 2021. Potential idiomatic expression (pie)-english: Corpus for classes of idioms. *arXiv preprint arXiv:2105.03280*.

Rodrigo Agerri. 2008. Metaphor in textual entailment. In *Coling 2008: Companion volume: Posters*, pages 3–6, Manchester, UK. Coling 2008 Organizing Committee.

Rodrigo Agerri, John Barnden, Mark Lee, and Alan Wallington. 2008. Textual entailment as an evaluation framework for metaphor resolution: A proposal. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 357–363. College Publications.

Yuri Bizzoni and Shalom Lappin. 2018. Predicting human metaphor paraphrase judgments with deep neural networks. In *Proceedings of the Workshop on Figurative Language Processing*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015b. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel R. Bowman, Jennimaria Palomaki, Livio Baldini Soares, and Emily Pitler. 2020. New protocols and negative results for textual entailment data collection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8203–8214, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021a. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021b. MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261, Online. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado. Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The Pragmatics behind Politics: Modelling Metaphor, Framing and Emotion in Political Discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488, Online. Association for Computational Linguistics.

Adith Iyer, Aditya Joshi, Sarvnaz Karimi, Ross Sparks, and Cecile Paris. 2019. Figurative usage detection of symptom words to improve personal health mention detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1142–1147, Florence, Italy. Association for Computational Linguistics.

Arthur M Jacobs. 2018. The Gutenberg English poetry corpus: exemplary quantitative narrative analyses. *Frontiers in Digital Humanities*, 5:5.

Charles Jochim, Francesca Bonin, Roy Bar-Haim, and Noam Slonim. 2018. SLIDE - a sentiment lexicon of common idioms. In *Proceedings of the Eleventh International Conference on Language Resources and*

Evaluation, Miyazaki, Japan. European Language Resources Association.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.

Ioannis Korkontzelos, Torsten Zesch, Fabio Massimo Zanzotto, and Chris Biemann. 2013. SemEval-2013 task 5: Evaluating phrasal semantics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 39–47, Atlanta, Georgia, USA. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press, Chicago and London.

Jens Lemmens, Ilia Markov, and Walter Daelemans. 2021. Improving hate speech type and target detection with hateful metaphor features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16, Online. Association for Computational Linguistics.

Vladimir Iosifovich Levenshtein. 1965. Binary codes capable of correcting deletions, insertions, and reversals. In *Doklady Akademii Nauk*, volume 163, pages 845–848. Russian Academy of Sciences.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and WordNet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Yixin Nie, Yicheng Wang, and Mohit Bansal. 2019. Analyzing compositionality-sensitivity of nli models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6867–6874.

Geoffrey Nunberg, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language*, (3):491–538.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191, New Orleans, Louisiana. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Ekaterina Shutova. 2013. Metaphor identification as interpretation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 276–285, Atlanta, Georgia, USA. Association for Computational Linguistics.

Ekaterina V Shutova. 2011. Computational approaches to figurative language. Technical report, Cambridge University.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins.

Anatol Stefanowitsch and Stefan Th. Gries. 2008. *Corpus-Based Approaches to Metaphor and Metonymy*. De Gruyter Mouton.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736, Online. Association for Computational Linguistics.

Chang Su, Shuman Huang, and Yijiang Chen. 2017. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Baltimore, Maryland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Tony Veale, Ekatarina Shutova, and Beata Beigman Klebanov. 2016. *Metaphor: A computational perspective*. Morgan and Claypool.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 3266—3280, Red Hook, NY, USA. Curran Associates Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shweta Yadav, Jainish Chauhan, Joy Prakash Sain, Krishnaprasad Thirunarayan, Amit Sheth, and Jeremiah Schumm. 2020. Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 696–709, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions*, pages 33–48, Online. Association for Computational Linguistics.

A Model Hyperparameters

We use a fixed set of hyperparameters for all NLI fine-tuning experiments: learning rate of $1e^{-5}$, batch size 32, and maximum input length of 128 tokens. The models are trained for 3 epochs. We used the HuggingFace implementation of the models (Wolf et al., 2020).

B Lexical Overlap

Previous research shows that NLI systems exploit cues based on lexical overlap, predicting entailment for overlapping sentences (McCoy et al., 2019; Nie et al., 2019). Our dataset consists mostly of pairs with high overlap: this could explain why the non-entailment sections are more difficult. We thus evaluate system predictions for our datasets as a function of lexical overlap. Figure 5 shows density-based histograms of the results, comparing overlap via Levenshtein distance (Levenshtein, 1965) for correctly and incorrectly classified pairs.

Our data contains higher overlap than the MNLI data, with the bulk of the density falling on minimally distant pairs. We also note a distinct difference between our entailment and non-entailment pairs: non-entailments contain extremely high overlap and are frequently misclassified in these cases where the distance is small, matching previous reports for NLI tasks: lexical overlap is a key artifact for entailment, and this reliance persists when classifying idiomatic pairs.

C Syntactic Flexibility Correlations

Figure 6 shows correlations between ICE scores (determined by frequency of occurences of a given IE outside of its normal form) and `roberta-base` model performance on that IE.

D Dataset Examples

Table 6 shows examples from each type of pair generation.

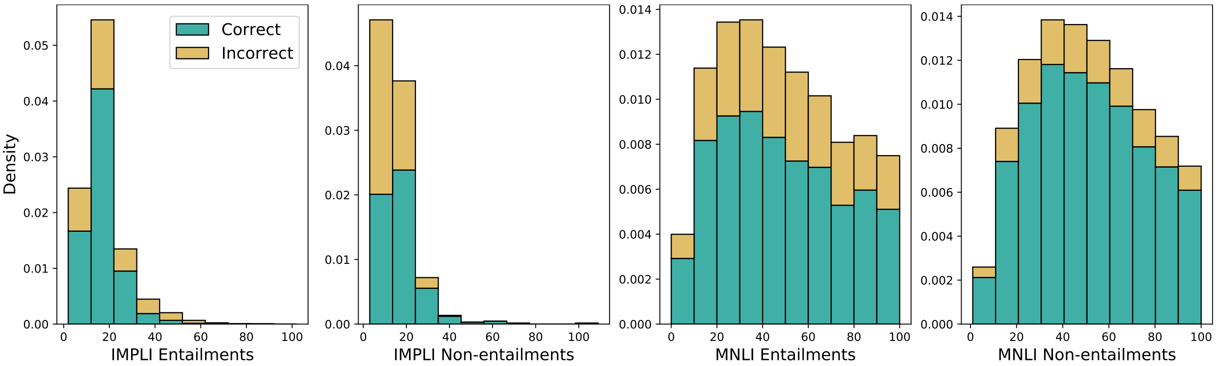

Figure 5: **R2: Lexical Overlap**. Classification performance by lexical overlap. The x-axis shows Levenshtein distance; the y-axis shows stacked density of correctly and incorrectly tagged pairs. The IMPLI non-entailments contain extremely high overlap, and are thus frequently misclassified as entailment.

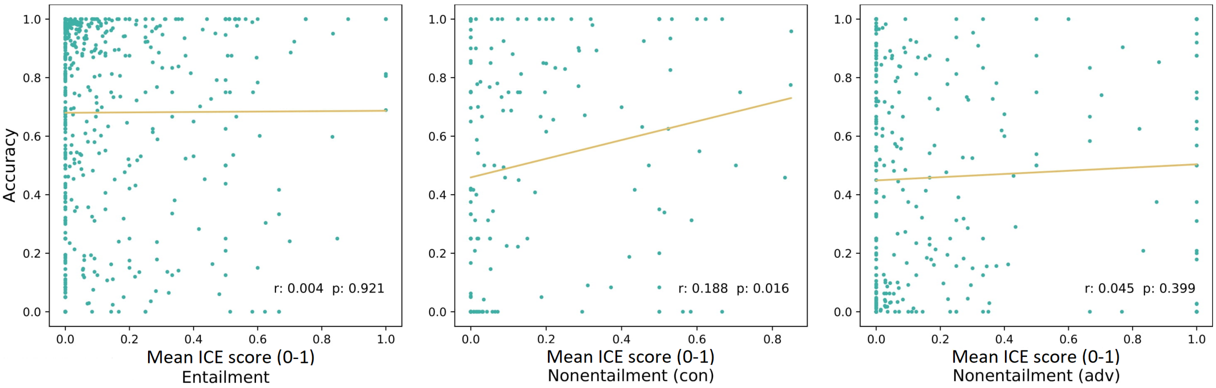

Figure 6: **R3: Syntactic Flexibility**. Performance of idiom types compared to their syntactic flexibility (based on ICE score defined in **R3**), with Spearman coefficient correlations r and significance values p. The middle figure is non-entailments based on replacement in literal context; the right is those based on adversarial definitions. Further right on the x-axis indicates greater flexibility.

Idioms
(→ S) Replace idiom used in figurative context with definition
BITTER BLOW: Beer sales are *feeling the pinch*. → BITTER BLOW: Beer sales are *suffering a hardship*.
I must *have a word* with them. → I must *speak privately* with them.
I’ve been knocked *out cold*. → I’ve been knocked *unconscious*.
(↛ S^l) Replace idiom used in literal context with definition
It would be good to roll *in hot water* all over. ↛ It would be good to roll *in a difficult situation* all over.
Pour in *the soup*. ↛ Pour in *trouble*.
There’s a marina down *in the docks*. ↛ There’s a marina down *under scrutiny*.
(↛ S^d) Replace idiom used in figurative context with adversarial definition
After *taking a bow*, the cast met Margaret backstage. ↛ After *apologizing*, the cast met Margaret backstage.
I’ve been knocked *out cold* ↛ I’ve been knocked *out into the cold air*.
It worked *like a charm*! ↛ It worked *poorly*!
(→ G) Hand written literal definition of idiom
How have you *weathered the storm*? → How have you *succeeded in getting through the difficult situation*?
It *breaks my heart* that his career has been ruined. → It *overwhelms me* that his career has been ruined.
Jamie rushed out *pissed off* and upset this afternoon. → Jamie rushed out *irritated* and upset this afternoon.
(↛ G^a) Manual replacement of key words in definition w/ antonyms
Alison *makes the grade* for Scotland ↛ Alison *fails* for Scotland.
I’ll *catch a cold* ↛ I’ll become *healthy*
It’s very much *swings and roundabouts* ↛ It’s very much *one-sided*.
(↛ G) Hand written non-entailed sentence
How have you *weathered the storm*? ↛ How have you *calmed the storm*?
Now Paul will *think twice*. ↛ Now Paul will *score twice*.
They *went to ground* somewhere in the area. ↛ They *went to party* somewhere in the area.
Metaphors
(→ S) Replace metaphoric construction with literal construction
Do not go and *blow* your *paycheck*. → Do not go and *waste* your *paycheck*.
My computer *battery died*. → My computer *battery lost all power*.
Competition is *dropping prices*. → Competition is *reducing prices*.
(→ G) Hand written literal paraphrase of metaphor
He *absorbed* the knowledge or beliefs of his tribe. → He *mentally assimilated* the knowledge or beliefs of his tribe.
Avon *treads* warily. → Avon proceeds warily.
All the *hearts of men were softened*. → All the *men were made kindler and gentler*.
(↛ G) Hand written non-entailed sentence
The gun kicked back into my shoulder. ↛ The mule kicked back into my shoulder.
This was conveniently *encapsulated* on the first try. ↛ This was conveniently *encapsulated* in the first battle.
On their tracks his eyes were *fastened*. ↛ On their tracks his *hands* were fastened.

Table 6: *Dataset Summary*: Overview of each entailment/non-entailment category in the IMPLI dataset.