IMPLI : Investing NLI Models' Performance on Figurative Language

Anonymous ACL submission

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 that require understanding of figurative languages such as idioms and metaphors remains understudied. We introduce the IMPLI (Idiomatic and Metaphoric Paired Language Inference) dataset consisting of over 25K semi-automatically generated and 1.5K hand-written English sentence pairs based on idiomatic and metaphoric phrases. We use IMPLI to evaluate NLI models based on RoBERTa fine-tuned on the MNLI dataset, and show that while they can reliably detect entailment relationship between figurative phrases with their literal definition, they perform poorly on examples where the phrases are designed to not entail the paired definition. This dataset suggests the limits of current NLI models with regard to understanding figurative language and provides 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 still struggle with multiword expressions in which individual words do not directly contribute to the sentiment (Sag et al., 2002). 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).

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.

Deep pretraining 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' representation 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., 2018a; Nie et al., 2020), but figurative language suffers from a lack of the necessary paired data, in which a literal sentence is linked to a corresponding figurative counterpart. 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 figurative 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 methods to create a large number of viable pairs (§3.1), as well as hand-written gold pairs (§3.2), crafted

[1] Dataset and all related resources (including Datasheet for Datasets document) are included in the supplementary materials.

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 involving 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 the representations they form for particular words and/or phrases will not reflect their typical 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, idioms, and other figures 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), but as metaphors are known to be extremely common and important to understanding figurative language, our resource builds 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 text (*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 text 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 these figurative areas. 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 into literal expressions. We expand on this methodology by instead having annotators only correct definitions for the idioms themselves, and using 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 20k silver sentences. We also expand beyond

Fig. Type	Ent.	Gold/silver	Description	Count
Idioms	$\rightarrow$	Silver	Replace idiom used in figurative context with definition	16652
	$\not\rightarrow$	Silver	Replace idiom used in literal context with definition	886
	$\not\rightarrow$	Silver	Replace idiom used in figurative context with adversarial definition	6116
	$\rightarrow$	Gold	Hand written literal definition of idiom	532
	$\not\rightarrow$	Gold	Manual replacement of key words in definition w/ antonyms	375
	$\not\rightarrow$	Gold	Hand written non-entailed sentence	254
Metaphors	$\rightarrow$	Silver	Replace metaphoric construction with literal construction	375
	$\rightarrow$	Gold	Hand written literal paraphrase of metaphor	388
	$\not\rightarrow$	Gold	Hand written non-entailed sentence	282

Table 2: **Dataset Summary**: Overview of each entailment/non-entailment category in the IMPLI dataset. $\rightarrow$ denotes entailments, $\not\rightarrow$ non-entailments. Note that the examples are simplified: some intermediate steps described in §3 are not shown above.

paraphrasing by incorporating both entailment and non-entailment pairs to enable evaluation via an NLI set-up.

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.

We focus on idioms and metaphor, using diverse methodology to build novel pairs with the goal of helping to fill the figurative language gap in diagnostic datasets. 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, built from both silver pairs (§3.1) and manually written sentences (§3.2). For our purposes, we follow McCoy et al. (2019) to conflate the neutral and contradiction categories into a non-entailment label, since the distinction between the two can often be unclear. We then label every pair as either entailment ($\rightarrow$) or non-entailment ($\not\rightarrow$). Due to the difficult nature of the task and to avoid issues with crowdsourcing (Bowman et al., 2020), we train linguistics graduate students as expert annotators. Table 2 contains a complete overview of the different entailment and non-entailment types collected (More detailed examples are also provided in Appendix D).

3.1 Silver pairs

First, we explore a method for generating silver pairs which uses 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*. $\rightarrow$ He's stuck in bed, which is his *bad luck*.

Conversely, in cases 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*. $\not\rightarrow$ Switzerland is famous for six cheeses, sometimes referred to as *bad luck*.

We now cover the implementation details for generating silver idiom and metaphor pairs.

3.1.1 Idioms

To build idiomatic pairs, we use three corpora that contain sentences containing idiomatic expressions

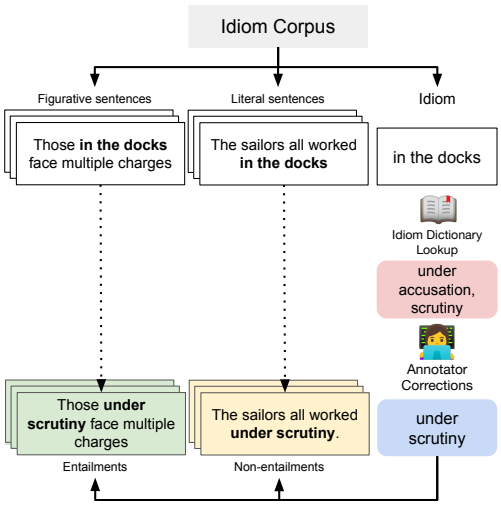

Figure 1: **Idiomatic definition replacement**. Pairs are generated using corrected dictionary definitions, substituted into figurative (left) and literal (center) sentences.

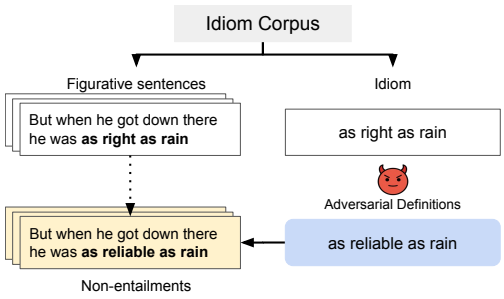

Figure 2: **Adversarial Pair Generation**. Non-entailing pairs are generated by replacing adversarial definitions into figurative contexts.

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

Table 3: Sample hand-written adversarial definitions.

(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 the corpus. Next, we 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 criteria wasn't 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). Replacing them into the original figurative contexts yields entailment relations, while replacing them into contexts where the phrase is meant literally then yields non-entailments.

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 take these adversarial definitions and replace them into the 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.

[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

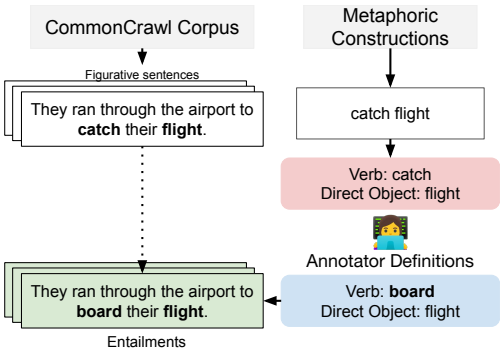

Figure 3: **Metaphor entailment generation**. Pairs are generated using annotator-defined literal translations substituted into metaphoric contexts.

1. circumstances *demand* → circumstances *require*
2. *drop* prices → *reduce* prices
3. *hairy* problem → *difficult* problem

These can then be replaced in a similar fashion to the idiom replacement: we start with the original figurative sentence, replace the ME with the literal replacements, and this results in an entailing pair with the metaphoric sentence entailing the literal.[4]

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[5], 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. For all silver methods, we employ syntactic pattern matching and regularization to ensure quality results (Appendix A), and we performed a manual evaluation of the pair quality (Appendix B).

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.

[4]Note this often works in both directions, but not always: some literal replacements don't necessarily entail the metaphors. Consider "Her husband abuses alcohol" → "Her husband drinks alcohol": the metaphor entails the literal, but not vice-versa.

[5]https://commoncrawl.org/

3.2 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, rephrasing, or adjusting the figurative component of the sentence. This yields gold standard paraphrases of 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. $\not\rightarrow$ 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. $\not\rightarrow$ You must adhere the rules to the wall.

3.3 Antonyms

Previous work in figurative language NLI 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; we explore this with regard to idioms using this dataset.

These various manual annotations provide a number of concrete benefits. First, they aren't restricted to individuals words or phrases (excluding antonyms): the figurative components can be

Model	MNLI	MNLI-MM	Idioms						Metaphors		
			$\rightarrow$ S	$\not\rightarrow$ S^l	$\not\rightarrow$ S^d	$\rightarrow$ G	$\not\rightarrow$ G^a	$\not\rightarrow$ G	$\rightarrow$ S	$\rightarrow$ G	$\not\rightarrow$ 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 4: **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.

rewritten freely, allowing for diverse, interesting pairs. Second, they are written by experts, ensuring higher quality than the automatic annotations, which may be prone to syntactic errors.

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 pretrained models perform on figurative entailments and non-entailments?*
- R2: *Does lexical overlap affect performance?*
- R3: *Does adding idiomatic pairs into the training data affect model performance?*
- R4: *Does the flexibility of idiomatic expressions affect model performance?*

R1: Pretrained Model Performance

We obtain standard NLI models by fine-tuning `roberta-base` and `roberta-large` models on the MNLI dataset (Williams et al., 2018b) (using the entailments as the possible class and all others as the negative class), and evaluate them on their original test sets as well as our **IMPLI** dataset.[6] To evaluate the 3-way classification models on the 2-way **IMPLI** dataset, we translate the predicted labels of *contradiction* and *neutral* into a *non-entailment* class. We report results in Table 5. Due to variance in neural model performance (Reimers and Gurevych, 2017), we average the results over 5 runs using different seeds.

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: they contain high lexical overlap, and in many cases similar semantic concepts to the literal contexts. The replacement into literal samples is more contrastive, with the idiomatic definition clashing more starkly with the original premise, making non-entailment predictions more likely. Consistent with Chakrabarty et al. (2021a)'s work in metaphors, we find non-entailment through antonym replacement is the easiest for idioms: the models can likely use the antonymic relationship as a marker for non-entailment, despite the high word overlap.

With regard to metaphors, the 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 pretrained models, as they are able to effectively capture the metaphoric relationships.

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: 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 4 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 re-

[6] Model hyperparameters found in Appendix C.

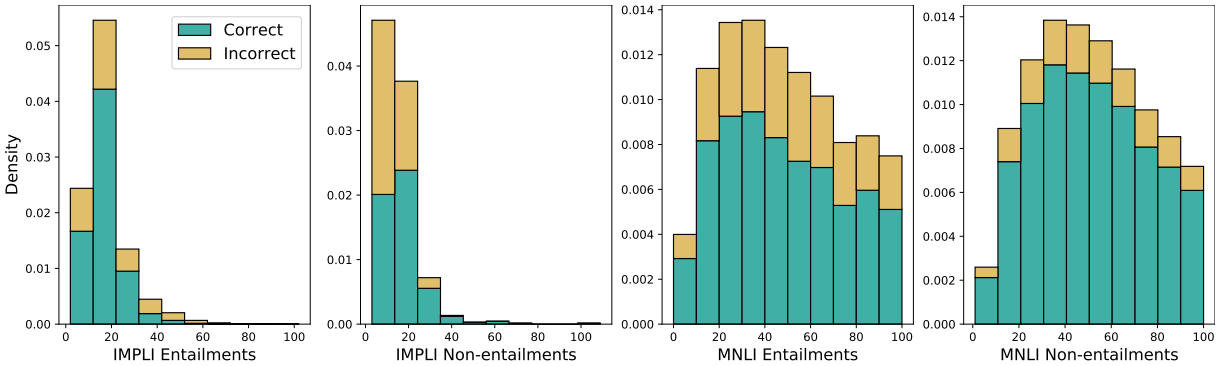

Figure 4: **R2: Lexical Overlap**. Classification performance by lexical overlap. The x axis shows Levenshtein distance; the y axis shows stacked density of correctly and incorrectly tagged pairs. The IMPLI non-entailments contain extremely high overlap, and are thus frequently misclassified as entailment.

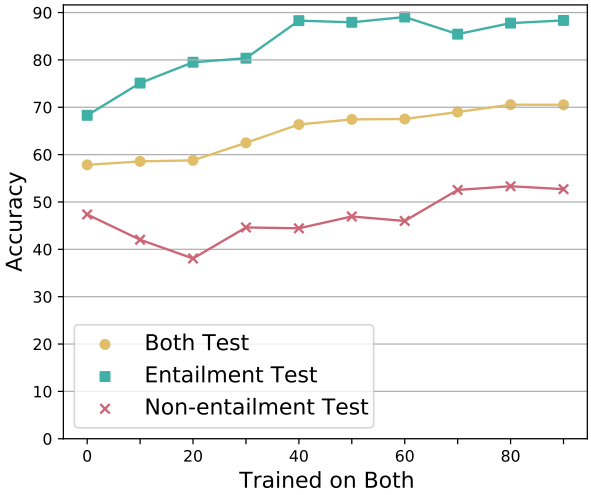

Figure 5: **R3: Training**. Performance of the `roberta-base` model as idiom types are added to the training data. Performance on improves slightly but remains difficult even when all training data is added.

ports for NLI tasks: lexical overlap is a key artifact for entailment, and this reliance persists when classifying idiomatic pairs.

R3: 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 only in the test data to assess whether the model can learn to correctly handle novel 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, leaving one out for testing, and incrementally incorporate this data into the original MNLI for training. 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.

The results shown in Figure 5 highlight some weaknesses with regard to idioms. Additional training data yields only small improvements: non-entailment relations remain exceedingly difficult, with performance capping out at only slightly better than chance. Additional training data may be somewhat effective in improving language models' idiomatic capabilities, but this is not sufficient to overcome difficulties from literal usages of idiomatic phrases and adversarial definitions, indicating that figurative language remains difficult for pretrained language models to learn to represent.

R4: 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 some divulging of information, and the beans to the information. Because this expression has individual semantic mappings to the figurative meaning for its constituents, they argue that they can be more syntactically flex-

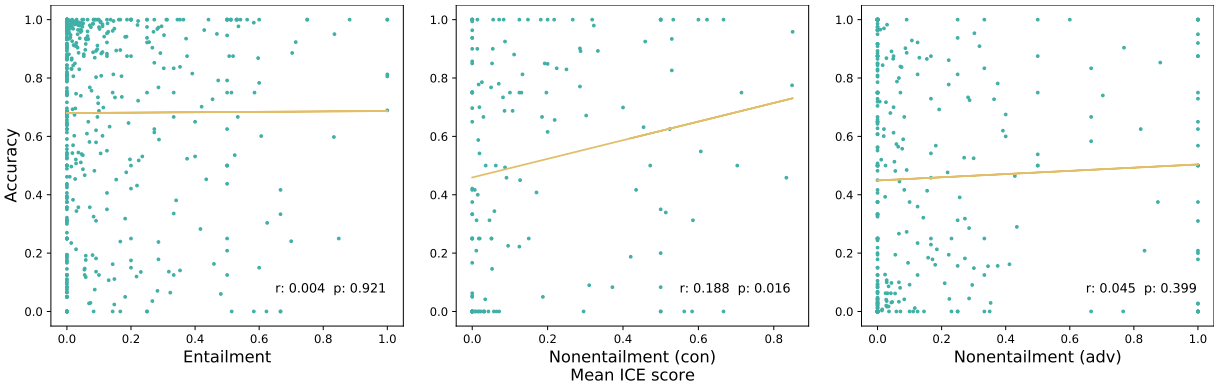

Figure 6: **R4: Syntactic Flexibility**. Performance of idiom types compared to their syntactic flexibility, with Spearman coefficient correlations. The middle figure is non-entailments based on replacement in literal context; the right is those based on adversarial definitions. Further right indicates greater flexibility.

ible. Compare to fixed expressions such as *kick the bucket*: no part of this expression maps directly to the figurative meaning of "die", and thus we expect less syntactic flexibility.

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 metric as the percentage of times a phrase occurs in our test data in a form that doesn't 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 plot the performance of the `roberta-base` model for each idiom type along with its ICE value.

Figure 6 shows the results. Interestingly, we see no correlation between ICE scores and performance for entailments, nor for non-entailments based on adversarial definitions. However, we do see a weak but significant correlation ($r = .188, p = 0.016$) with non-entailments from literal contexts: the model actually performs better when the phrases are more flexible, contrary to our hypothesis that ICEs would be more difficult.

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, consisting of 24k silver and 1.8k gold figurative/literal pairs in English, which we then use to evaluate NLI models' capabilities on figurative language. We show that while standard NLI models handle entailment admirably, and metaphoric expressions are relatively easily, non-entailment idiomatic relationships are much 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 NLG typically relies on training or fine-tuning models with paired sentences, this data could also be a valuable resource for training figurative NLG 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 (EMNLP)*, 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 (LREC 2018)*, Miyazaki, Japan. European Language Resources Association.

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.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

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, Ł ukasz 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. *arXiv preprint 1905.00537*.

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. 2018a. 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. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018b. 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.

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 (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

A Postprocessing

We employ syntactic postprocessing to overcome a number of hurdles. First, in many cases, phrases used idiomatically often carry a different syntactic usage than when they are used literally. Consider the idiom *out of this world*, and the given definition "wonderful":

> **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 meaning functions syntactically as an adjective, thus the definition "wonderful" produces a grammatically incorrect sentence when replacing the original. Second, the phrases in their literal usage often don't form full constituents. This is due to the string-matching approach of the original datasets. Many literal usages of these phrases are thus incompatible with the defined replacement. Consider the phrases *to die for* and *in the raw* below:

- 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. We required 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.

A.1 Metaphor Guidelines

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 doesn't actually contribute to the metaphor, but rather the entire expression is necessary. Relatedly, we found that replacing individual parts of MEs is often incapable of fully removing the metaphoric meaning. There are frequent examples of personification and changing position on a scale where metaphoric elements may persist. We iterated over possible solutions to

	$\rightarrow$ Idioms	$\not\rightarrow$ Idioms	$\rightarrow$ Met.
Syntax	.92	.80	.83
Entailments	.88	.90	.97

Table 5: **Valid pairs**. Percentage of valid pairs, syntactically and with regard to the intended entailments, of automatic data generation, $n = 100$ per category.

circumvent these issues; we eventually decided to simply skip instances for which a replacement does not yield a feasible literal interpretation.

B Verifying Automatic Pair Quality

The automatic methods used to recast idiomatic and metaphoric data into usable NLI pairs require additional quality control: in order to evaluate the quality of the recasted generation, we additionally validated whether the automatically generated pairs were in fact syntactically valid sentences, and contained the appropriate entailment relation. For this task, each annotator was given 100 samples for each of the silver generations (idiomatic entailments, idiomatic non-entailments, and metaphoric entailments). They were asked whether the automatically generated pair was syntactically valid, and what the entailment relation between the two sentences was. We adjudicated disagreements to determine the final percentage of valid pairs.

Scores for these metrics range between .80 and .97. Idiomatic non-entailments had poor syntax, due to the difficulties described in §3.1. Metaphoric entailments also were somewhat weak, but on inspection we find most of these syntactic errors to be minor issues regarding verb conjugations.

C 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.

D Dataset Examples

Table 6 shows examples from each type of pair generation.

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

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