Figurative Language in Recognizing Textual Entailment

Tuhin Chakrabarty[*1], **Debanjan Ghosh**[* 3], **Adam Poliak**[2,4] **and Smaranda Muresan**[1,4]
[1]Department of Computer Science, Columbia University
[2]Department of Computer Science, Barnard College
[3]Educational Testing Service, [4]Data Science Institute, Columbia University
{tuhin.chakr, smara}@cs.columbia.edu,
dghosh@ets.org, apoliak@barnard.edu

Abstract

We introduce a collection of recognizing textual entailment (RTE) datasets focused on figurative language. We leverage five existing datasets annotated for a variety of figurative language – simile, metaphor, and irony – and frame them into over 12,500 RTE examples.We evaluate how well state-of-the-art models trained on popular RTE datasets capture different aspects of figurative language. Our results and analyses indicate that these models might not sufficiently capture figurative language, struggling to perform pragmatic inference and reasoning about world knowledge. Ultimately, our datasets provide a challenging testbed for evaluating RTE models.

1 Introduction

Figurative language is ubiquitous in many forms of discourse from novels, poems, and films, to scientific literature and social media conversations (Ghosh, 2018). It is often used to convey intimacy (Gerrig and Gibbs Jr, 1988), humour (Roberts and Kreuz, 1994), intense emotions (Fussell and Moss, 1998), or veiled politeness (Jorgensen, 1996). Despite its ubiquity, figurative language remains "a bottleneck in automatic text understanding" (Shutova, 2011).

Recognizing Textual Entailment (RTE), the task of identifying whether one sentence (context) likely entails another (hypothesis), is often used as a proxy to evaluate how well Natural Language Processing (NLP) systems understand natural language (Cooper et al., 1996; Dagan et al., 2006; Bowman et al., 2015). Figurative language is defined as any figure of speech which depends on a non-literal meaning of some or all of the words used. Thus, understanding figurative language can be framed as an RTE task (figurative language expression vs. intended meaning), where the figurative language expression is the *context* and the intended meaning is the *hypothesis* in an RTE framework (See examples in Table 1).

Simile	► I start to prowl across the room like a **tightrope walker on dental floss**. I start to prowl across the room *recklessly*.	✗
	► They had shut him in a basement that looked like a **freight elevator**. They had shut him in a basement that looked *dangerously claustrophobic*.	✓
Metaphor	► He **weathered** the costs for the accident. He *avoided* the costs for the accident.	✗
	► The bus **bolted** down the road. The bus *paced* down the road.	✓
Irony	► Made $174 this month, gonna buy a yacht! I don't make much money.	✗
	► Fans seem restless, gee, don't understand them. Fans seem restless - don't know the reason behind it.	✓

Table 1: Example RTE pairs focused on similes, metaphors, and irony that RoBERTa *incorrectly* labels. ► indicates a context and the following sentence is its corresponding hypothesis. ✓ and ✗ respectively indicate that the context entails, or does not entail the hypothesis. **Bold** text represent simile and metaphors and *Italic* represent their entail/not entail interpretations (top two rows).

We investigate how suitable are state-of-the-art RTE models trained on current RTE datasets to capture figurative language. We focus on three specific types of figurative language: similes, metaphors, and irony. Similes evoke comparisons between two seemingly different objects, metaphors expand the imagination beyond the literal narrative, and irony conveys the opposite of what is said.

We leverage five existing datasets annotated for these types of figurative language to create over

[*]Equal Contribution.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3354–3361
August 1–6, 2021.

12,500 RTE examples that require understanding or identifying these phenomena. We evaluate how well standard neural RTE models capture these aspects of figurative language. Our results demonstrate that, although, systems trained on a popular RTE dataset may capture some aspects of various types of figurative language, they fail on cases where the interpretation relies on pragmatic inference and reasoning about world knowledge. We release the code and the data. [1]

2 Related Work

We follow recent work that test for an expanded range of inference patterns in RTE systems (Bernardy and Chatzikyriakidis, 2019) by evaluating how well RTE models capture specific linguistic phenomena, such as pragmatic inferences (Jeretic et al., 2020), veridicality (Ross and Pavlick, 2019), and others (Pavlick and Callison-Burch, 2016; White et al., 2017; Dasgupta et al., 2018; Naik et al., 2018; Glockner et al., 2018; Kim et al., 2019; Kober et al., 2019; Richardson et al., 2020; Yanaka et al., 2020; Vashishtha et al., 2020; Poliak, 2020).

We are not the first to explore figurative language in RTE. Agerri (2008) analyze examples in the Pascal RTE-1 (Dagan et al., 2006) and RTE-2 (Bar-Haim et al., 2006) datasets that require understanding metaphors and Agerri et al. (2008) present an approach for RTE systems to process metaphors. Poliak et al. (2018)'s diverse collection of RTE datasets includes examples based on figurative language, but focuses only on identifying puns.

3 Dataset Creation

We create RTE test sets that focus on similes, metaphors, and irony. We provide further background for these types of figurative language and describe the methods used for creating these test sets. Table 2 reports the final test sets' statistics.

3.1 Simile

Comparisons are inherent linguistic devices that express the likeness of two entities, concepts, or ideas. When used figuratively, comparisons are called similes. Similes are used to spark the reader's imagination by making descriptions more emphatic or vivid (Paul et al., 1970). Similes use a common PROPERTY to compare two concepts often referred to as the TOPIC (the logical subject) and the VEHICLE (the logical object of comparison). For example, in the simile "Love is like an unicorn", love (TOPIC) is compared to a unicorn (VEHICLE), portraying the implicit property "rare". Recently Chakrabarty et al. (2020) released a test set of 150 literal sentences from subreddits r/WritingPrompts and r/Funny, each aligned with two human-written paraphrases with similes that retain the original meaning.

Data		Total	E	NE
Simile		600	300	300
Metaphor		613	307	306
Irony Meaning	$SIGN_{2000}$	2,000	133	1867
	S_{im}-H_{int}	4,762	-	4,762
Irony Intention		4,601	2,212	2,389

Table 2: Dataset statistics and class distribution, *Entailment* (E) and *Not-Entailment* (NE) for each type of figurative language.

To create our RTE test set that focuses on similes, we treat these simile-literal aligned sentences as entailed context-hypothesis pairs. Given a literal input, "They had shut him in a basement that looked **dangerously claustrophobic**", an expert annotator re-framed it as "They had shut him in a basement that looked *like a freight elevator*".[2] We create Not-Entailed examples by flipping the literal verb/property with their respective antonyms and use the original (Literal, Simile) pairs as Entailed. For instance, in the case of an existing context-hypothesis pair expressing *Entailment* - "Hitler skittered off like **an enthusiastic sloth**" → "Hitler skittered off *slowly*" - we alter "slowly" to "fast" to make it a pair of *Not-Entailment* (NE) instance.

3.2 Metaphor

Metaphors express deep feelings and complex attitudes (Veale et al., 2016). Understanding metaphors requires comprehending abstract concepts and making connections between seemingly unrelated ideas to appropriately deviate from literal meaning (Gutierrez et al., 2016; Mohammad et al., 2016; Kintsch and Bowles, 2002; Glucksberg, 1998).When generating metaphoric paraphrases, Chakrabarty et al. (2021) create a diverse test set of 150 literal sentences curated from different domains and genres and asked two expert annotators to create metaphorical sentences, resulting in a total

[1] https://github.com/tuhinjubcse/Figurative-NLI

[2] Note, such re-framing task (content generation task) does not involve assigning a label to a text fragment, thus, computing inter-annotator agreement is not applicable here.

Genre	PairID	Example
Slate	143311e	▶ Praise from a stranger is **like a glass of water** served at a restaurant in: You drink it warily, if at all, fearing it may be tainted Praise from someone you do not know can be taken lightly
Fiction	60838c	▶ The stars are no more like the sun than the glow of my cigarette is **like a forest fire**. The sun is comparable to the stars because they are the same.
Telephone	99298c	▶ But uh still I I question the ability of some of the teachers to uh really do a **bang-up job** and yet others i know are just wonderful All teachers sucks

Table 3: Examples from MNLI that include figurative language. ▶ indicates a context and the following line is its corresponding hypothesis.

of 300 metaphorical examples. The expert annotators re-framed the literal sentences independently by replacing the literal verb with a metaphorical verb. For instance, an expert reframed the literal sentence "The tax cut will help the economy" to "The tax cut will **fertilize** the economy".

Since the most frequent type of metaphor is expressed by verbs (Martin, 2006; Steen, 2010) these literal and metaphorical paraphrases differ only by the verb they use. In an RTE framework, we treat these metaphorical-literal pairs as entailed context-hypothesis examples. To create Not-Entailed examples, we generate hypotheses by manually swapping the literal verb in the entailed hypothesis with its antonym. Note that for both simile and metaphor, automatic substitution using available lexicons is problematic as it often leads to ungrammatical sentences. Manually replacing the words with its antonym guarantees a high quality test set. We use antonyms to create Not-Entailed examples for Simile and Metaphors which contain both Neutral and Contradiction classes. Such lexical replacement using antonyms would clearly lead to higher quality contradiction example creation. On the contrary, creating neutral examples by lexical perturbation is challenging and if not done properly, it can lead to grammatical errors or incoherent sentences.

3.3 Irony

When using irony, speakers usually mean the opposite of what they say (Sperber and Wilson, 1981; Dews et al., 2007). We develop different test sets focusing on whether the RTE models should *understand the conveyed meaning* of ironic examples or should *identify the speaker's ironic intent* (i.e., identify if an utterance is ironic or not) given the hypothesis that the speaker was ironic.

Understanding Ironic Meaning (IMeaning) Peled and Reichart (2017) used skilled annotators to create a parallel dataset between tweets with verbal irony and their non-ironic rephrasings (15K pairs). Annotators also had the option to copy the original tweet or just to paraphrase it, in case the ironic intent is not easy to identify. Likewise, Ghosh et al. (2020) released a parallel dataset of speakers' ironic messages (S_{im}) and hearers' interpretations (H_{int}) of the speaker's intended meaning. This dataset (S_{im}-H_{int}) contains 4,761 ironic-literal pairs. We use both datasets in our experiments and henceforth denote them as $SIGN$ and S_{im}-H_{int}, respectively. For both datasets, the original *ironic* messages are treated as the contexts and the *intended* meanings are the hypotheses. However, all RTE contexts do not contradict their corresponding hypotheses. For instance, in case of Peled and Reichart (2017), the authors allowed annotators to not rephrase the ironic sentences with their opposite *intended meanings* (in case the sarcastic or ironic intent was not clear). Thus, for evaluation purposes (see Table 4), we annotated a subset of 2,000 random pairs from $SIGN$ and evaluated the RTE models on that subset (denoted as $SIGN_{2000}$ henceforth). Around 93% of the RTE pairs in $SIGN_{2000}$ are Not-Entailed examples and 100% of RTE pairs in S_{im}-H_{int} are Not-Entailed examples.

Recognizing Ironic Intent (IIntent) We leverage additional ironic examples from Van Hee et al. (2018). Following Poliak et al. (2018)'s method for recasting annotations for puns and sentiment, we use *templates* to generate contexts (a) and hypotheses (b). We use all the ironic tweets ($training$ and $test$) released by Van Hee et al. (2018) to generate 4,598 RTE pairs. Akin to Poliak et al. (2018), we

Testset / Model	Simile	Metaphor	IMeaning		IIntent
			$sm - im$	$SIGN_{2000}$	
NBoW	51.17	54.81	86.37	71.50	**61.72**
InferSent	55.01	65.75	71.62	68.84	11.72
RoBERTa-large	**85.47**	**88.09**	**94.76**	**93.42**	52.81

Table 4: Accuracy of different models on our datasets focusing on similes, metaphors, and irony.

replace *Name* with names sampled from a distribution of names based on the US census data.[3]. The templates are a) *Name* tweeted that *tweet*, b) *Name* was ironic.

4 Experimental Setup

MNLI (Williams et al., 2018) is one of the widely used large-scale corpora that contains instances of figurative language (Table 3). Following recent work, we evaluate RTE models trained on MNLI (Williams et al., 2018) using three standard neural models: bag of words (NBoW) model, InferSent (Conneau et al., 2017), and RoBERTa-large (Liu et al., 2019). In NBoW, word embeddings for contexts and hypotheses are averaged separately, and their concatenation is passed to a logistic regression softmax classifier. InferSent encodes the context and hypotheses independently using a BiLSTM, then their sentence representations are fed to a MLP.[4] For RoBERTa, we use the model fine-tuned on MNLI from the Transformer's library (Wolf et al., 2020). We expect models trained on MNLI to capture some forms of figurative language that often appear in works of fictions, conversations, speeches, and magazines like Slate. Table 3 illustrates a few examples from MNLI that include figurative language

5 Results and Discussions

Table 4 reports models' accuracy on our figurative language RTE datasets. We observe that for similes, metaphors and irony meaning, RoBERTa-large drastically outperforms the other two models. For Irony datasets, NBoW outperforms InferSent. While all models perform poorly on IIntent, InferSent's very low accuracy stands out. The low performances might be due to the templatic nature of this recast dataset which might be very different from the MNLI training data.[5] We now turn to an in-depth analysis of RoBERTa's performance across these datasets.

Ironic Meaning. RoBERTa-large attains over 90% accuracy on the two datasets focused on ironic meaning. When analyzing these examples, a vast majority of the hypotheses in both datasets use lexical antonyms ("flattering" ↔ "disgusting) or negation ("is great" ↔ "is not great") to represent the intended meaning. Thus, the presence of antonyms might be enough for RoBERTa to correctly predict that the hypothesis is not-entailed by the context.

However, this does not hold true for hypotheses where the intended meanings were represented via more complex rephrasing. Ghosh et al. (2020) conducted a thorough study of the *linguistic strategies* that annotators have used for the rephrasing tasks. They presented a linguistically motivated typology of the strategies (e.g., "Lexical and phrasal antonyms", "Negation", "Weakening the intensity of sentiment", "Interrogative to Declarative Transformation", "Counterfactual Desiderative Constructions", and "Pragmatic Inference") and empirically validated the strategies over the $SIGN$ and S_{im}-H_{int} datasets.[6] During our analysis, we observe that for the vast majority of cases where RoBERTa predicts incorrectly, the examples contain Rhetorical Questions ("nice having finals on birthday?" ↔ "do not like finals ..."), pragmatic inferences ("Made $174 this month ... a yacht!" ↔ "I don't make much money"), or desiderative constructions of *[I wish] (that)* ("glad you related the news" ↔ "[I wish] that you have told me sooner". We also observe that RoBERTa-large's predictions are regularly incorrect when the ironic messages contain certain irony markers (Ghosh and Muresan, 2018), such as metaphor ("shoe smell like bed of roses" ↔ "smells bad"), alternate spelling where the speaker frequently overstate the magnitude of an ironic event ("dancing in heels is grrrrreat" ↔ "... hurts your feet") or hashtags that are composed of multiword expressions that capture the irony ("god bless you ... #notinthemood).

[3] http://www.ssa.gov/oact/babynames/names.zip

[4] Both NBoW and InferSent use 300D Glove embeddings (Pennington et al., 2014).

[5] We leave further analysis of this issue for future work.

[6] https://github.com/debanjanghosh/interpreting_verbal_irony

			Gold	Pred
Simile	▶	Your guardian angel is just a little too much like a **nerd at a comic convention**. Your guardian angel is just a little too *enthusiastic*	✓	✗
	▶	Growing up, people always thought you were like a **social pariah**. Growing up, people always thought you were *ordinary*	✗	✓
	▶	They all agree the books are good reads, but they are like **pseudo science fiction**. They all agree the books are good reads, but they are *too unrealistic*.	✓	✗
Metaphor	▶	The smell of smoke **carpeted** on the delinquent. The smell of smoke *took off* on the delinquent	✗	✓
	▶	As they strike the ground, they are **effaced**. As they strike the ground, they are *remembered*	✗	✓
	▶	The avalanche **polvarized** anything standing in its way. The avalanche *protected* anything standing in its way.	✗	✓
Irony	▶	Life was never been perfect and would never be. Life has never been perfect and would never be.	✓	✗
	▶	The highlight of my day figuring out how to make contact sheets . . . such a boring life. My entire day was occupied in making contact sheets in design such a waste.	✓	✗
	▶	Gotta read 70ish+ pages today #great #mysundayfunday #thisshouldbefun. I have to read 70ish+ pages today. This is bad.	✗	✓

Table 5: Examples from our Simile, Metaphor, and Irony datasets where Roberta-large fine-tuned on MNLI fails to classify the sentence pairs correctly. Gold and Pred means the true label and the predicted label respectively. ▶ indicates a context and the following sentence is its corresponding hypothesis. ✓ and ✗ respectively indicate that the context entails, or does not entail the hypothesis.

Simile. Likewise, for the simile dataset, we notice that RoBERTa-large often fails to reason with implicit knowledge about the physical and visual world knowledge (Table 5). This is inline with Weir et al. (2020)'s finding that transformer-based contextual language models poorly capture knowledge grounded in visual perceptions. For example, RoBERTa-large incorrectly predicts that the context "You wake one morning to find your entire family lying like **gray slabs of cement**" does not entail the hypothesis "You wake one morning to find your entire family lying *unconscious*". Nevertheless, RoBERTa-large correctly predicts that, "my eyes teared up . . . turning like a **ripening tomato**" entails "my eyes teared up . . . face *turning red*". We hypothesize that here RoBERTa-large was able to identify the association between "ripening tomato" and "red" that resulted in the correct prediction.

Metaphor. We notice RoBERTa-large makes wrong predictions when it encounters *unconventional* metaphors (Table 5). Metaphors are deemed unconventional depending on "how well-worn or how deeply entrenched a metaphor is in everyday use by ordinary people for everyday purposes" (Gelo and Mergenthaler, 2012). For instance, for a unconventional (metaphoric, literal) pair, "night sky **flurried** with the massive bombardment" → "night sky *doused* with the massive bombardment" (i.e., "flurried" ↔ "doused") the model fails. On the contrary, the model correctly predicts the following conventional (metaphoric, literal) pair - "sudden fame **kindled** her ego" → "*. . . increased* her ego" (i.e., "kindled" ↔ "increased").

6 Conclusion

To understand the figurative language inference capabilities of RTE models, we introduce datasets adapted from existing corpora focusing on similes, metaphors, and irony. By testing models trained on MNLI, we find that while the RoBERTa-large model is able to capture some aspects of figurative language, it fails when the interpretation requires word knowledge and pragmatic inferences. We hope this work will spark additional interest in the research community to incorporate and test for figurative language in their NLU systems.

7 Ethical Considerations

We leverage freely available open source datasets and software tools to create RTE datasets that involve similes, metaphors, and irony. We are granted the rights to further annotate and distribute the existing datasets as part of our RTE setup. This research is exempt from institutional review boards since we do not study human subjects and all social media data used is publicly available.

References

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.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, and Bernardo Magnini. 2006. The second pascal recognising textual entailment challenge.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2019. What kind of natural language inference are nlp systems learning: Is this enough?

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6455–6469, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. 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.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework.

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. Springer.

I. Dasgupta, D. Guo, A. Stuhlmüller, S. J. Gershman, and N. D. Goodman. 2018. Evaluating Compositionality in Sentence Embeddings. *ArXiv e-prints*.

Shelly Dews, Joan Kaplan, and Ellen Winner. 2007. Why not say it directly? the social functions of irony. *Irony in language and thought*, pages 297–318.

Susan R Fussell and Mallie M Moss. 1998. Figurative language in emotional communication. *Social and cognitive approaches to interpersonal communication*, pages 113–141.

Omar Carlo Gioacchino Gelo and Erhard Mergenthaler. 2012. Unconventional metaphors and emotional-cognitive regulation in a metacognitive interpersonal therapy. *Psychotherapy Research*, 22(2):159–175.

Richard J Gerrig and Raymond W Gibbs Jr. 1988. Beyond the lexicon: Creativity in language production. *Metaphor and Symbol*, 3(3):1–19.

Debanjan Ghosh. 2018. *An Empirical Study of Verbal Irony*. Ph.D. thesis, Rutgers, The State University of New Jersey.

Debanjan Ghosh and Smaranda Muresan. 2018. "with 1 follower i must be awesome :p". exploring the role of irony markers in irony recognition. *Proceedings of ICWSM*.

Debanjan Ghosh, Elena Musi, and Smaranda Muresan. 2020. Interpreting verbal irony: Linguistic strategies and the connection to theType of semantic incongruity. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 82–93, New York, New York. Association for Computational Linguistics.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking nli systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.

Sam Glucksberg. 1998. Understanding metaphors. *Current Directions in Psychological Science*, 7(2):39–43.

E Dario Gutierrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 183–193.

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705, Online. Association for Computational Linguistics.

Julia Jorgensen. 1996. The functions of sarcastic irony in speech. *Journal of pragmatics*, 26(5):613–634.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics.

Walter Kintsch and Anita R Bowles. 2002. Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and symbol*, 17(4):249–262.

Thomas Kober, Sander Bijl de Vroe, and Mark Steedman. 2019. Temporal and aspectual entailment. In *Proceedings of the 13th International Conference on Computational Semantics-Long Papers*, pages 103–119.

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*.

James H Martin. 2006. A corpus-based analysis of context effects on metaphor comprehension.

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.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2340–2353, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Anthony M Paul, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 1970. Figurative language. In *Philosophy & Rhetoric*, pages 225–248.

Ellie Pavlick and Chris Callison-Burch. 2016. Most "babies" are "little" and most "problems" are "huge": Compositional entailment in adjective-nouns. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2164–2173. Association for Computational Linguistics.

Lotem Peled and Roi Reichart. 2017. Sarcasm sign: Interpreting sarcasm with sentiment based monolingual machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1690–1700, Vancouver, Canada. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Adam Poliak. 2020. A survey on recognizing textual entailment as an nlp evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J. Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting diverse natural language inference problems for sentence representation evaluation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 337–340, Brussels, Belgium. Association for Computational Linguistics.

Kyle Richardson, Hai Na Hu, Lawrence S. Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *AAAI*, volume abs/1909.07521.

Richard M Roberts and Roger J Kreuz. 1994. Why do people use figurative language? *Psychological science*, 5(3):159–163.

Alexis Ross and Ellie Pavlick. 2019. How well do NLI models capture verb veridicality? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2230–2240, Hong Kong, China. Association for Computational Linguistics.

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

Dan Sperber and Deirdre Wilson. 1981. Irony and the use-mention distinction. *Philosophy*, 3:143–184.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. Semeval-2018 task 3: Irony detection in english tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50.

Siddharth Vashishtha, Adam Poliak, Yash Kumar Lal, Benjamin Van Durme, and Aaron Steven White. 2020. Temporal reasoning in natural language inference.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. Metaphor: A computational perspective. *Synthesis Lectures on Human Language Technologies*, 9(1):1–160.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. In *42nd Annual Virtual Meeting of the Cognitive Science Society (CogSci)*.

Aaron Steven White, Pushpendre Rastogi, Kevin Duh, and Benjamin Van Durme. 2017. Inference is everything: Recasting semantic resources into a unified evaluation framework. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 996–1005.

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. 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.

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *ACL*.