Metaphor and Entailment

Looking at metaphors through the lens of textual entailment.

Artemis Panagopoulou

A THESIS

in

Computer and Information Science

Presented to the Faculties of the University of Pennsylvania in Partial
Fulfillment of the Requirements for the Master of Science in Engineering

2020

Professor Mitch Marcus
Supervisor of Thesis Signature

Professor Mayur Naik
Graduate Group Chairperson Signature

Abstract

Metaphors are very intriguing elements of human language that are surprisingly prevalent in our everyday communications. Humans are pretty good at understanding metaphors, even if it is the first time they encounter them [33]. Empirical studies indicate that 20% of our daily language use is metaphorical[82]. Naturally, the ubiquity of metaphors draw the attention of psychologists who have shown that the human brain processes conventional metaphors at the same speed as literal language[41, 44, 51, 50].

Nevertheless, the computational linguistics literature consistently treats metaphors as a separate domain to literal language. Earlier work has shown that traditional pipelines do not perform well on metaphoric datasets[63, 1]. Synchronously, the literature on computational understanding of metaphors has largely focused on developing metaphor detection systems coupled with interpretation systems targeted solely on metaphors[73]. This tendency has presented across various aspects of the field, such as the purposeful exclusion of figurative language from large scale datasets[1]. This study investigates the potential of constructing systems that can jointly handle metaphoric and literal sentences by leveraging the newfound capabilities of deep learning systems.

We narrow the scope of the report, following earlier work[63, 1], to evaluate deep learning systems fine-tuned on the task of textual entailment (TE). We argue that TE is a task naturally suited to the interpretation of metaphoric language. We show that TE systems can improve significantly in metaphoric performance by being fine tuned on a small dataset consisting of metaphoric premises. Even though the improvement in performance on metaphors is typically accompanied by a drop in performance on the original dataset we note that auto-regressive models seem to show a smaller drop in performance on literal examples compared to other types of models.

"Logic is a very elegant tool,and we've got a lot of mileage out of it for two thousand years or so. The trouble is, you know, when you apply it to crabs and porpoises, and butterflies and habit formation you know, to all those pretty things logic won't quite do...because that whole fabric of living things is not put together by logic. You see when you get circular trains of causation, as you always do in the living world, the use of logic will make you walk into paradoxes."
"So what do they use instead?"
"Metaphor."
"Metaphor?"
"Yes, metaphor. That's how the whole fabric of mental interconnections holds together. Metaphor is right at the bottom of being alive."

Fritjof Capra, Uncommon Wisdom: Conversations with remarkable people (1988)
Bantam, New York [Interview with Gregory Bateson, page 76-77]

Contents

List of Figures

List of Tables

1 Introduction

Metaphors are very intriguing elements of human language that are surprisingly prevalent in our everyday communications. Humans are pretty good at understanding metaphors, even if it is the first time they encounter them [33]. In fact, empirical studies indicate that 20% of our daily language use is metaphorical[82]. Naturally, the ubiquity of metaphors draw the attention of psychologists who showed that the human brain processes conventional metaphors at the same speed as literal language[41, 44, 51, 50].

Nevertheless, the computational linguistics literature consistently treats metaphors as a separate domain to literal language. Earlier work has shown that traditional pipelines do not perform well on metaphoric datasets[63, 1]. Synchronously, the literature on computational understanding of metaphors has largely focused on developing metaphor detection systems coupled with interpretation systems targeted solely on metaphors[73]. This tendency has presented across various aspects of the field, such as the purposeful exclusion of figurative language from large scale datasets[1]. This study investigates the potential of constructing systems that can jointly handle metaphoric and literal sentences by leveraging the newfound capabilities of deep learning systems.

We narrow the scope of the report, following earlier work[63, 1], to evaluate deep learning systems fine-tuned on the task of textual entailment (TE). We argue that TE is a task naturally suited to the interpretation of metaphoric language. We show that TE systems can improve significantly in metaphoric performance by being fine tuned on a small dataset with metaphoric premises. Even though the improvement in performance on metaphors is typically accompanied by a drop in performance on the original dataset we note that auto-regressive models seem to show a smaller drop in performance on literal examples compared to other types of models.

1.1 Motivation

TE is a task naturally suited for metaphor interpretation. As discussed in section 2.2.1, most prominent theories of metaphor can be re-formulated in the context of TE. Metaphors can either be viewed as entailing a comparison, or a literal paraphrase, or a series of class inclusion statements. Interestingly, George Lakoff and Mark Johnson explicitly use the term *metaphorical entailments* to characterize the underlying conceptual mappings that, according to their theory, make effortless metaphoric interpretation possible[53][1].

Therefore, it is unsurprising that previous work has examined the efficacy of earlier TE models on metaphoric inputs. [1] evaluated textual entailment systems on a small set of metaphors[2] and discovered that their performance on the metaphorical examples was significantly reduced compared to the overall accuracy on the dataset. A similar study[63] was repeated with a larger evaluation dataset, only to reach a similar conclusion that the state of the art systems of the time were not suitable for metaphoric inputs. Interestingly, the authors of these two studies had conflicting views on how to deal with that difficulty. In [1], the author concludes that we should strive to improve the performance of textual entailment systems on metaphors to increase their accuracy, whereas in [63], the authors assert that it would be most productive to develop an accurate metaphor identification systems and treat metaphoric and literal language separately.

In this thesis, we perform a similar analysis on current state of the art textual entailment methods to measure any potential improvement in terms of metaphorical language resolution. We hypothesize that the advent of heavily parameterized deep learning models, which are already trained on both metaphoric and literal sentences ((see section 4.2) may be better suited than earlier machine learning models to interpret metaphors. Consequently, this study aims to investigate the potential of deep learning to support the development of joint models for literal and metaphorical language through the tuning and evaluation of such pre-trained language models on metaphoric data.

1.2 Contribution

This research project makes the following contributions:

- A reliable dataset of 1397 annotated entailment pairs with metaphorical premises. This is the largest such dataset available to the best of our knowledge. The only comparable (but still

[1] This theory will be further elaborated upon in section 17
[2] 10 examples

smaller) such dataset is the one used in [63] but it is not publicly available. Apart from the entailment task label, the dataset contains annotations about metaphor type to allow for finer analysis.

- An application of state of the art metaphor detection[58] on MultiNLI dataset[101], a popular TE dataset used to train deep learning models. This investigation aims to verify the claim that TE models are already trained on metaphors as well as evaluate the performance of metaphor detection systems on crowdsourced examples.
- An evaluation of state of the art deep learning natural language systems on the new dataset with and without fine-tuning on metaphors.

1.3 Roadmap

This study will be organized as follows: Section 2 provides the linguistic background relevant to this project. Specifically, section 2.2 discusses theories of metaphor with the aim to render entailment as a natural fit to the task of metaphor interpretation. The same section discusses theories of metaphor identification that have influenced the computational linguistics literature[3], as well as an outline of types of metaphors that are relevant to the current study. Section 2.1 introduces the notion of logical entailment in natural language. Section 3 discusses computational efforts on metaphors and textual entailment. Specifically, section 3.1 discusses existing metaphoric datasets and methods for metaphor detection and interpretation. Section 3.2 discusses natural language entailment datasets and computational approaches with a focus on deep learning systems as they are most relevant to this study[4]. Section 4.1 introduces the metaphoric TE dataset constructed as part of this study. Section 4.2 contains a report on applying state of the art metaphor detection models[58] on MultiNLI[101] and shows that current deep learning TE systems are already trained and tested on metaphors. Section 5 presents experiments with the newly constructed metaphoric

[3] In this project special attention is placed on metaphor identification with a twofold aim: first, to evaluate the extend to which large scale natural language corpora already contain metaphors. Second, to show that such systems are currently not effective enough to serve their purpose of acting as an arbitrator to identify whether a sentence should be processed through a literal or metaphoric-tuned pipeline.

[4] For a longer review of natural language entailment systems see [20]

TE dataset. We include hypothesis-only baselines[72] for the new data in 5.1 and report results of finetuning models in section 5.2. Finally, section 6 discusses future approaches in metaphor processing informed by the results of this study.

2 Linguistic Background

This section provides an overview of the linguistic background relevant to this study. Section 2.1 presents the task of textual entailment. Section 2.2.1 presents prominent theories of metaphor, and argues that all share the property of being readily framed within the context of textual entailment. Section 2.2.2 discusses theories of metaphor identification which have informed computational such systems discussed section 3.1.2. Section 2.2.3 distinguishes between different types of metaphors relevant to this project.

2.1 Entailment

Textual entailment has been a central task for natural language understanding that has witnessed significant improvements in recent years with the advent of deep learning models. The task of textual entailment takes as input two natural language sentences A, B and outputs one of three values: 1. *entailment* in the case that B must be true if A is true, 2. *contradiction* in the case that B cannot be true if A is true, and 3. *neutral* in the case that B could be true, but A does not logically entail it. Note that the sentence A is called a *premise* and sentence B a *hypothesis*.

A simple example can make this task very clear. Consider the premise 2 and the three associated hypotheses: 4, 5, and 6. We can see that hypothesis 4 is *entailed* in the premise 2 since it follows logically that if 2 is true, then 4 must be true. On the contrary, if 2 is true, then 5 cannot be true, and thus this is an example of *contradiction*. Finally, hypothesis 6 could be true if 2 is true, but it could also be false; in this case we consider the sentences to have a *neutral* relation.

Premise: (1)

All men are mortal and Socrates is a man. (2)

Hypotheses: (3)

Socrates is mortal (4)

Socrates is immortal. (5)

Socrates has a beard. (6)

Part of the beauty and difficulty in dealing with natural language is the fact that it is often indeterminate; especially when considered out of its context. This is important to note, because unlike the aforementioned examples, there are many cases in which TE labeling is not as clear-cut. Below I provide such an example which is extracted from the MultiNLI[101] development corpus through the identification of examples with highest annotation disagreement[5].

Premise: You want to punch the button and go (7)

Hypothesis: You don't want to push the button lightly, but rather punch it hard (8)

In this case there is no clear interpretation: three of the five annotators concluded that this example is neutral. However, one annotator considered this to be an entailment and another a contradiction. This is because the phrase "punch the button and go" could be interpreted as either indicating an intense energetic movement, or a light and fast movement.

The next section investigates the theoretical underpinnings of metaphors, and attempts to draw a parallel between metaphors and TE.

2.2 Metaphors

The word metaphor etymologically stems from the Greek work "μεταφερείν" which means ”to transfer", alluding to the core of metaphoric vernacular, which is the transfer of attributes from one

[5] There were more examples with the same exact score of disagreement. This one was selected at random from that pool.

domain (the source) to another (the target).

For example, consider the following famous metaphor

$$\text{Love is blind} \tag{9}$$

In metaphor 9 the target domain is blindness, which is an attribute for animate beings, and the source domain is love. So the target domain is applied to the source domain to elicit the meaning that one can love another regardless of physical qualities.

Metaphors are of interdisciplinary interest, thus making it difficult to provide a single definition that captures the concept in its entirety. A series of different definitions of metaphors have been popularized across disciplines and they are often induced from associated theories of metaphors.

For example, Oxford Learner's Dictionaries defines metaphor as *"a word or phrase used to describe somebody/something else, in a way that is different from its normal use, in order to show that the two things have the same qualities and to make the description more powerful, for example She has a heart of stone"*[6]. As we will see later in this section, the definition implicitly upholds the comparison theory of metaphors.

Notice that the aforementioned definition identifies metaphors at a word level. In this study it will be important to define metaphors at a sentence level. When is a *sentence* metaphoric? According to the philosopher Max Black *"in calling this former sentence a metaphor, we are implying that there is one word that is being used metaphorically"*[11]. We will use this definition when referring to a metaphoric sentence throughout this study.

In the rest of the section we discuss the theoretical underpinnings of metaphoric language and suggest that they can all be framed within the context of textual entailment.

2.2.1 Theories of Metaphor

Metaphor is an indispensable and ubiquitous component of discourse. In fact, a study [22] found that 19% of the words in a broad sample of texts, ranging from poetry to science, were metaphors.

[6] https://www.oxfordlearnersdictionaries.com/us/definition/english/metaphor

It is not surprising that this figure of speech has attracted the attention of thinkers since ancient times. As we will see in this section, famous ancient philosophers such as Plato and Aristotle had expressed (differing) opinions on metaphors. The quest of understanding metaphors has continued to this day, with a series of contesting views presented in the literature, some of which can trace back to the ancient times.

This section presents the most prevalent such theories that have arose from a series of different disciplines, such as linguistics, philosophy, and psychology. The major theories discussed in this section are: Subsitution Theory, Comparison Theory, Class Inclusion Theory, Interaction Theory, and Conceptual Mapping Theory. We argue that despite their differences all these theories can be interpreted in the context of TE.

2.2.1.1 Substitution Theory According to substitution theory, metaphors are a purely decorative figure of speech that can be replaced by a literal phrase that expresses the exact same meaning.

For example, consider the metaphor

$$\text{The classroom was a zoo.} \tag{10}$$

It is highly unlikely that the phrase is uttered literally. Instead, the speaker intended the sentence to be interpreted metaphorically. According to the substitution view, the speaker could have used a literal expression but chose to use a metaphoric one[11]. In this example, the speaker could have instead opted to use a literal adjective to characterize the classroom, for example, *"The classroom was loud"*.

It is very easy to see how substitution theory can be reformulated in the context of TE: it proposes a view where metaphors entail literal phrases[7]. According to this reformulation, metaphors can be thought as premises and the literal substitution as an entailed hypothesis. This structure is very closely related to the experimental setup proposed in section 5.

[7] It is worth noting that paraphrase is a more suitable context for substitution theory of metaphor, however, even paraphrase has often been studied closely to TE[4]

One of the early proponents of the substitution theory is Richard Whately, who discusses metaphors in his book *Elements of Rhetoric* (1828). Whately defines a metaphor as *a word substituted for another on account of the Resemblance or Analogy between their significations*. He further defined the notions *Resemblance* and *Analogy* as follows: *"Resemblance [is] the direct resemblance between the objects themselves in question"* and *"Analogy which is the resemblance of ratios a similarity of the relations they between the objects themselves in question."* In other words, according to Whately, metaphors can occur in two major ways: the first one is by direct resemblance between objects, such as when we compare waves to mountains. The second one, is the analogy between certain aspects of the objects, such as when we talk about the *"light of reason"*. In this case, reason does not physically exhibit luminosity, but instead the metaphor alludes to the clarity that can arise from logical reasoning[98].

In fact, proponents of substitution theory have often considered metaphors as a lesser, less precise choice of vernacular. Famously, Whately stated in his *Elements of Rhetoric* that metaphors often consist a deviation from the *"the plain and strictly appropriate style"*. It is interesting to note that this view was partly shared by Plato who believed that metaphors cannot be of more than heuristic value[14][8].

Despite its popularity in the 19th and early 20th centuries, substitution theory has accrued a significant amount criticism. Most famously, in the mid 1950s Professor Max Black argued that it is not possible to paraphrase all metaphors into *"literal statements"* without *"a loss of cognitive content"*[11].

2.2.1.2 Comparison Theory Metaphors have attracted the attention of thinkers since ancient times. In chapter 22 of the Poetics [37], Aristotle states

"The greatest thing by far is to be a master of metaphor; it is the one thing that cannot be learnt from others; and it is also a sign of genius, since a good metaphor implies an intuitive perception of the similarity in the dissimilar."

Aristotle's view was later formalized by other linguist and philosophers as the Comparison The-

[8] Despite his use of metaphor in his writings, such as the analogy of the sun in the Republic[42]

ory, or Classical Theory of metaphors. According to this theory any metaphor can be substituted with a comparison (or a simile[9]). Recall the example 10; according to comparison theory, we could re-frame this utterance into a comparison as follows

$$\text{The classroom was like a zoo (with respect to loudness).} \tag{11}$$

Often Comparison Theory has been considered a special case of the Substitution Theory, presented in the previous subsection, since it supports that any metaphor can be *substituted* with a literal comparison. It is thus clear that Comparison Theory can be reformulated in the context of TE in the same way as Substitution Theory: The metaphor is a premise entailing the literal comparison-based hypothesis.

One of the most prominent counter arguments about Comparison Theory is the fact that the conversion of a metaphor to a comparison is ambiguous, and sometimes, as Black states, *borders upon vacuity*[11]. The reason is that given a metaphor, such as 10 it is not clear what attribute is being compared. It could be loudness, as suggested in 11, but it could also be unruliness, competitiveness, and so on.

A more modern reformulation of comparison theory of metaphor has been provided by Diedre Gentner called Structure Mapping Theory[31]. This award winning theory[10] states that there can be more than one mapping that is referred to in a particular metaphor. In a way, Gentner removed the *"with respect to"* component of Comparison Theory, claiming that a metaphor can represent more of a structural mapping containing a series of similarities. For example, the metaphor 10 could be interpreted as a mapping between the classroom and a zoo in many levels: in terms of organization, loudness, diversity, and so on. Gentner further provides a systematic way of blending human attribution of analogy to a particular metaphor based on the similarity between the source and target domains.

Despite the eloquence of Gentner's revived Comparison theory of metaphors, there is still sig-

[9] Simile is a figure of speech that makes an explicit comparison signified with a preposition such as 'like', 'similar to', etc.

[10] Gentner was awarded the 2016 David E. Rumelhart Prize for Contributions to the Theoretical Foundations of Human Cognition

nificant doubt cast on its validity from empirical evidence that indicate that humans understand similes very differently to metaphors. For example, a 2006 study[33] studied the interpretations of metaphors and their equivalent analogies of 16 undergraduate students who were native speakers of English and found significant variation in interpretation. In particular, they noted that generally similes were interpreted faster than metaphors when the metaphor was more general, whereas metaphors were translated faster when accompanied by a more specific attribute, such as an adjective.

Another study[34] identified differences in comparisons and metaphors. The study compared literal and metaphoric comparisons by considering the meaningfulness of the original and reversed phrases (an example that elucidates this distinction can be seen in table 1). They found that even though literal comparisons were equally meaningful both in the original and the reversed formulation, this was not the case for metaphoric comparisons.

Type	Literal	Metaphors
Original	Yams are like potatoes.	Alcohol is like a crutch.
Reversed	Potatoes are like yam.	Crutch is like alcohol.

Table 1: Examples of Class-Inclusion Theory Empirical Evaluation

Such variations in interpretation have been cited[73] as a potential counter argument to the comparison view, showing at the very least that metaphors and comparisons cannot be thought as interchangeable.

2.2.1.3 Class Inclusion Theory According to class inclusion theory, metaphors are not implicit comparisons, but instead, class-inclusion assertions.

This theory is best understood through an example

A guitar is a fretted musical instrument. (12)

A laouto is like a guitar. (13)

My children are monsters. (14)

The phrase 12 is a literal class-inclusion sentence that categorizes a guitar as a fretted musical instrument. The second phrase 13, is again literal, and suggests that a laouto is similar to a guitar but it does not belong to the class of guitar. Note the similarity between this type of sentences and similes; they both emphasize similarities that are undoubtedly accompanied by differences.

The last phrase 14 is metaphoric and according to the class inclusion theory is a class-inclusion assertion that categorizes *my children* to belong to the class of *monsters* where *monsters* are used as a prototypical member of a generic class that represents *rowdy, disobedient, and mean* individuals.

Notice how this theory contrasts metaphors and similes. The class inclusion theory holds that metaphors are class-inclusion assertions with a dual reference: the first reference is to a literal sense and the second to the more abstract and generic class[21].

It is not difficult to observe that the Class Inclusion theory can naturally be reformulated in the context of TE: class inclusion statements can be thought as entailed hypotheses of the premise, which is the metaphor in question.

Critics of the theory suggest that it does not provide an explanation on how to interpret metaphors, as it does not provide a methodology on how the abstract classes are formed[53].

2.2.1.4 Interaction Theory According to the interaction theory of metaphor there exists a bidirectional relationship between the word used metaphorically and its surrounding context.

This theory was put forth by Max Black in 1962 as a response to the inability of the substitution and comparison theories to capture all types of metaphors. His work was informed from the writings of I.A. Richards who proposed an interactive relation between the terms used metaphorically (whose attributes are borrowed) which he called *vehicles* and the terms to which the attributes are ascribed, which he called *tenors*[78]. For example, in the metaphor “Love is blind” 9, *love* is the tenor, and *”blind”* is the vehicle.

Black vehemently holds that *“[a] metaphorical statement is not a substitute for a formal comparison or any other kind of literal statement, but has its own distinctive capabilities and achievements”*[11].

Black formulated his account of metaphor around what he called a *logical grammar*. He postulates this grammar for metaphor as the interactive relationship between the *focus* and the *frame* of the metaphor. The *focus* of the metaphor is the word used metaphorically, and the *frame* is the words surrounding it.

An ancient anger exploded in his heart (15)

Consider the example 15: in this case the *focus* of the metaphor is the word *exploded* and the *frame* the subject *an ancient anger* and the prepositional phrase *in his heart*. Black maintains that the original thoughts and meanings of the focus and the frame interact to generate a new meaning for the focus which is not its literal meaning, but is also not the same as any literal substitution[5]. In Black's words the focus' meaning *"is not quite its meaning in literal uses, nor quite the meaning which any literal substitute would have"*[11]. Black calls the process through which this new meaning is engendered the *system of associated commonplaces*[11, 5] which are the associations evoked to the listener when she imagines the focus and the frame, while the un-shared ideas are suppressed. In this example, the listener would imagine the common ideas/concepts between anger and explosion, such as intensity, pain, catastrophe, and leave out the un-shared ones, such as fire, smoke, death, etc.

Black's formulation of interaction theory, rejects the idea supported by the comparison view which claims that there exist pre-existing similarities between the source and target domains. Instead those similarities are evoked from the frame and the listener's background knowledge[40].

Unlike earlier theories, it is not obvious how interaction theory can be interpreted within the context of TE. In order to do so we need to place more attention on the *system of associated commonplaces*. .

Jake is a wolf. (16)

Consider the example 16 borrowed from Black's writings[11]. In this metaphor the frame is

‘John is’ and the focus is ‘a wolf’. In figure 1[11] we delineate the associated commonplaces with respect to the frame in set A, and the associated commonplaces with respect to the focus in set B. The intersection of these two sets comprises of the associations imposed on the focus through the frame. We can therefore think of a metaphor as entailing the intersection of associations which can be formulated as hypotheses of the form ‘The focus is X’, where X represents an association. For example, the metaphor 16 is a premise that entails the hypotheses ‘Jake is hairy’, ‘Jake is short tempered’, ‘Jake is active at night’, and so on.

There are issues that arise within this interpretation. At first glance it seems that this interpretation is more closely related to class inclusion theory, since there does not seem to be an interaction between the focus and the frame. However, this interaction is implied in the generation and calibration of the sets of commonplaces. The frame and the focus interact until they converge in a stable intersection. Class inclusion on the other hand, assumes a predefined set of attributes characterized by varying scales of abstractions associated with each word.

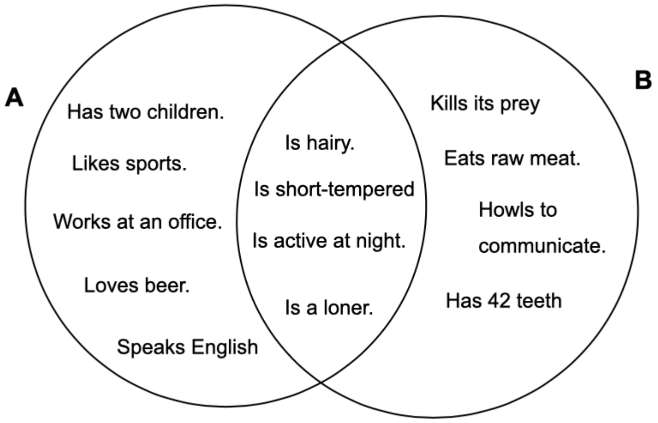


Figure 1: Schematic Explanation of Associated Commonplaces

2.2.1.5 Conceptual Mapping Theory This theory, which was introduced by George P. Lakoff and Mark Johnson[53], postulates that there exist fixed conceptual mappings between different domains that allow the generation of metaphors through adaptation in different contexts. A concep-

[11] The schematic is inspired by the figure in [5].

tual mapping can be defined as a mapping between a source and target domain that translates to "TARGET DOMAIN IS SOURCE DOMAIN".

We explain this theory by borrowing a famous example from Lakoff and Johnson's book "Metaphors We Live By" (2008) [53]

Conceptual Mapping: ARGUMENT IS WAR (17)

Manifestation 1: Your claims are indefensible. (18)

Manifestation 2: If you use that strategy, he'll wipe you out. (19)

Conceptual mapping theory rejects Black's notion of meaning extension, and instead supports a cognitive re-conceptualization view. Consider the conceptual mapping 17 and its manifestations 18, 19. We can see that in those manifestations the concept of an argument is re-conceptualized into a war, which involves an opponent as well as actions such as shooting and killing. Mapping those properties of war to an argument domain allows for the manifestations of the mapping to be interpreted, as well as the potential generation of novel manifestations.

This theory does not require reformulation to be placed in the context of TE, since it was formulated as such. Lakoff termed conceptual mappings as *metaphorical entailments* to highlight the fact that the relation between a metaphors and their conceptual mappings is one of entailment[53]. It is interesting to note that this formulation of metaphoric theory in the context of TE differs to the approach taken in this study. In this theory the entailed hypotheses are also metaphoric in some sense. This observation could inspire the application of highly effective deep learning TE models in the interpretation of metaphors directly. Nevertheless, this observation is beyond the scope of this study and is left for future investigation.

Despite the traction that this theory accumulated, a variety of empirical studies seem to contradict it. For example, a study [45] showed that there does not seem to be a need to extract underlying conceptual mappings to understand conventional metaphorical expressions through a comparison of reading times compared to literal expressions. However, this was not the case for non-conventional metaphors[12].

Another critic of the theory addresses the lack of clarity in the case where there is more than one conceptual mapping involved in interpreting a metaphoric expression[64].

2.2.2 Identifying Metaphors

Figurative language is notoriously subjective[97]. The lack of uniform criteria for metaphor identification impede evaluation of theoretical claims and computational models on metaphoric language[81]. It is thus important for the community to agree on a systematic way of identifying metaphors. This subsection discusses different attempts to standardize metaphor identification procedures. These formulations have directly informed the development of computational models on metaphor detection (see 3.1.2).

2.2.2.1 Violation of Selectional Preference According to this view, metaphors can be identified by the fact that they elicit some sense of incongruity in meaning when attempted to be interpreted literally. In other words, this theory highlights the differences between the source and target domain. This view was notably put forth by Walker Percy in his essay *Metaphor as a Mistake* (1958).

Percy summarizes the problem of metaphor in philosophy and linguistics as follows: *Metaphor has scandalized philosophers, including both scholastics and semioticists, because it seems to be wrong: it asserts an identity between two different things. And it is wrongest when it is most beautiful.* [69]. An interpretation of Percy's thesis show him sympathetic towards more cognitive/-conceptual views on metaphors. He believes that this ”error” from which metaphor arises contains meaning that no theory of meaning could capture. He holds that metaphor cannot be simply interpreted as a decoration akin to what substitution and comparison theories propose, but it cannot also just be left as an open interpretation highly congruent to individual subjective worldview interpretation, since humans tend to agree in their interpretations of metaphor. As will be discussed in a later section (see 3.1), Selectional Preference has significantly influenced the computational view on metaphors, and in particular metaphor identification[36]. In fact, it inspired one of the first

[12] The distinction between conventional and non conventional metaphors is discussed in a later subsection.

computational models on metaphor developed by Yorick Wilks in 1975[99].

Selectional preference has also been employed for various tasks spanning from Semantic Role Labeling[105] to sense disambiguation[60].

2.2.2.2 Metaphor Identification Procedure (MIP) MIP was originally proposed by Gerard Steen in 2002[87] and was further formalized by the Pragglejaz Group in 2007[13] with the aim to align the community's intuition about metaphors. Their paper[35] delineates the procedure as follows[14]:

1. Read the entire text–discourse to establish a general understanding of the meaning.
2. Determine the lexical units in the text–discourse
3. (a) For each lexical unit in the text, establish its meaning in context, that is, how it applies to an entity, relation, or attribute in the situation evoked by the text (contextual meaning). Take into account what comes before and after the lexical unit.

   (b) For each lexical unit, determine if it has a more basic contemporary meaning in other contexts than the one in the given context. For our purposes, basic meanings tend to be

   - More concrete [what they evoke is easier to imagine, see, hear, feel, smell, and taste];
   - Related to bodily action;
   - More precise (as opposed to vague);
   - Historically older;
   - Basic meanings are not necessarily the most frequent meanings of the lexical unit.

   (c) If the lexical unit has a more basic current–contemporary meaning in other contexts than the given context, decide whether the contextual meaning contrasts with the basic meaning but can be understood in comparison with it.
4. If yes, mark the lexical unit as metaphorical.

[13] Steen was part of the group.
[14] verbatim.

For a full example of the procedure one can refer to the original paper.

It is interesting to note that there are difficulties associated with applying the procedure without adjustments in languages other than English, as this study[93] on Lithuanian showed.

2.2.3 Types of Metaphors

In this subsection we discuss different types of metaphors.

2.2.3.1 Conventionality We first discuss what is meant by metaphor conventionality; conventionality is a graded property that expresses *"how well-worn or how deeply entrenched a metaphor is in everyday use by ordinary people for everyday purposes"*[47]. On the one extreme of the conventionality continuum, we have metaphors that have been integrated effortlessly in everyday vernacular. For example, the understanding of time as a resource[15] results in the conventional metaphor "Time is running out". This concept is closely related to the interpretation of metaphors as conceptual mappings, as discussed in an earlier subsection.

Non-conventional metaphors, on the other hand, are original and are often found in contexts such as poetry, literature, rhetoric, and so on[30]. Consider for example the first verse of Emily Dickinson's poem *"'Hope' is the thing with feathers"*[25] where she compares hope to a bird.

> ”Hope” is the thing with feathers -
> That perches in the soul -
> And sings the tune without the words -
> And never stops - at all -
>
> Emily Dickinson (1830–1886)

Different theories have been proposed on how non-conventional metaphors are generated. George Lakoff and Mark Turner postulated that non-conventional metaphors are created through the modification of an already existing conceptual conventional metaphor[91] whereas Zoltan Kövecses supports that they can be completely novel[48].

Psycholinguists have investigated differences in the time it takes for conventional and non-conventional metaphors to be interpreted by studying Event Related Protential (ERP) through

[15] Example from Glossary of linguistic terms

electroencephalography (EEG). Studies have shown that non-conventional metaphors take slightly longer to be interpreted than conventional metaphors[51, 50].

The notion of metaphor conventionality is relevant to this study because the metaphoric dataset composed for evaluation consists of metaphors that fall on the conventional side of the spectrum. In addition, the metaphoric examples identified in existing entailment datasets also fall in this category.

2.2.3.2 Contracted and Extended Metaphors Metaphors can be distinguished as contracted or extended. On the one hand, contracted metaphors are those whose effect is limited to the phrase or sentence boundary in which they appear. This is the type of metaphors this study is concerned with since it would be irrational to require an entailment model, which operates on a sentence level, to comprehend a metaphor that is not fully interpretable within the sentence it appears.

Extended metaphors, on the other hand, are metaphors that can span through longer discourse fragments, such as Dickinson's poem presented in an earlier section. Extended metaphors typically expand on a conceptual mapping; for example, Dickinson's extended metaphor revolves around the conceptual mapping HOPE IS A BIRD[73].

2.2.3.3 Linguistic Types of Metaphors Metaphors are categorized depending on their different linguistic features. We can distinguish between lexical and multiword metaphors. Lexical metaphors are single-word-level metaphors, whereas multiword comprise of two or more words. An example of a multiword metaphor can be found in 20. In this example, the subject is compared to be drowning in the sea to refer to their sadness that pertains through every aspect of their lives.

He was drowning in a sea of grief. (20)

Lexical metaphors can be categorized further into Nominal, Subject-Verb-Object, Adjective-Noun, and Adverb-Verb. We consider each of these types in turn.

Nominal Metaphors: A metaphor is nominal when the target domain is directly compared to the source domain through the use of a copular verb. For example, the metaphor 10 is a nominal metaphor.

Subject-Verb-Object Metaphors: These metaphors resolve around the use of a verb metaphorically and unlike nominal metaphors, their comparison is indirect. For example, in 21 the verb “shoot” is used metaphorically, employing the underlying conceptual mapping ARGUMENT IS WAR.

He shot down all my arguments (21)

Adjective Noun Metaphor: In this metaphor type, an adjective is used metaphorically to characterize a noun. For example, in 22 the adjective “black” is used metaphorically to characterize the noun mood.

She was in a black mood. (22)

Adverb-Verb Metaphors: This type of metaphor occurs when an adverb is used metaphorically. For example, in 23 the adverb ‘fluidly’ is used metaphorically to characterize her speech. This type of metaphor has not received as much attention in the computational linguistics literature as the other types[73].

She speaks fluidly. (23)

3 Previous Work

3.1 Metaphors

In this section we consider the computational background on metaphors. We discuss available corpora, as well as methods for detection and interpretation of metaphors. The in depth investigation of existing metaphoric datasets exposes the need for the dataset presented in section 4. Moreover, the detailed exploration of methods of metaphor detection aims to inform the selection of a metaphor detection model to explore the metaphoricity of MultiNLI[101] in section 4.2.

3.1.1 Datasets

A series of datasets on metaphors have been developed over the years. A summary of the relevant information of the datasets discussed is presented in table 2.

One of the first datasets was TroFi [8, 9] which was developed using a semi-supervised method. The dataset contains a total of 6445 sentences, 3746 of which where annotated by the self-supervised model with high confidence. The corpus contains 2145 non-literal and 1592 literal sentences. The semi supervised system uses a combination of verb-noun clustering and active learning, namely human intervention at each iteration to correct errors in the algorithm. The model achieves a reported 64.91% F1-score, when evaluated against human annotations of 25 verbs.

The Master Metaphor List (MML)[49] is another dataset of nominal metaphors. The dataset was compiled from the homonymous list of metaphors collected by George Lakoff, Jane Espenson, and Adele Goldberg August in 1989[52].

Another, very popular dataset on metaphors is the VU Amsterdam Metaphor Corpus (VUAMC)[86]. The corpus includes randomly sampled excerpts from a variety of genres and employs a collaborative adaptation of MIP to annotate their metaphoricity at a word level. The corpus has high annotation agreement 0.84[27] and consists of 187570 lexical units/115 distinct fragments. The corpus contains only metaphors, coupled with annotations regarding to whether the metaphor is clear or borderline, direct or indirect, and it further includes personification tags. The dataset consists of 1728 short sentences/nominal phrases, 939 of which are literal and 789 metaphorical.

Hovy et. all[39] released a set of 3872 sentences manually annotated on metaphoricity that were used to train a metaphor detection model based on dependency tree kernels.

One of the datasets we used in this study was Tsvetkov's[90]. The dataset includes 200 English SVO sentences and 200 English sentences containing a noun characterized by an adverb. From each section (SVO,AV) half sentences are metaphorical and the other half literal. The dataset is accompanied by a manually derived Russian translation of the dataset.

The next dataset we consider was not created for computational processes, but instead to conduct a psycholinguistics study[89]. It consists of 150 modal sentences, 75 of which are nominal metaphors and the other 75 literal.

Moh-X[62] is a dataset of 342 sentences annotated by metaphoricity and emotionality. It was used to evaluate the extend to which metaphors evoke stronger emotional responses compared to their literal counterparts.

More relevant to this study, Yuri Bizzoni and Shalom Lappin[10] collect a dataset of 200 metaphors, with associated sentences and paraphrase grades. We use this dataset to recast the dataset used in this study for the evaluation of state of the art entailment models. More information about this dataset is provided in the section that outlines the data used in this study.

More recently a metaphor dataset was crowdsourced from Twitter[106] using a weakly annotated model that proposed potential metaphoric tweets, and used human annotators to verify them. This methodology resulted in a set of of 2,500 manually annotated tweets in English with over 0.8 inter-annotation agreement. The data was focused around emotional tweets, and those concerning Brexit.

It is clear that the available datasets on metaphors treat metaphor as separate from daily vernacular. They are targeted to the task of metaphor identification, rather than blending metaphor within pre-existing tasks in natural language understanding despite the ubiquitousness of metaphor in daily language. This study attempts to break that pattern by investigating the incorporation of metaphors within general purpose TE systems.

Name	Year	Type	Size	Genre	Special Annotations
TroFi[8, 9]	2006	SVO	6445 sentences	News	Organized by verb
MLL[49]	2007	NO	1728 phrases 939 literal 789 metaphorical	-	-
VUAMC[86]	2010	MW	187570 words 115 fragments	academic conversations fiction news text	Clear/Borderline Direct/Indirect Personification
Hovy[39]	2013	MW	3872 sentences	-	-
Tsvetkov[90]	2014	SVO+AN	200 SVO 200 AV	-	Metaphor Type Russian
Thibodeau[89]	2015	NO	75 Metaphor 75 Literal	-	-
Moh-X[62]	2016	MW	171 Metaphors 171 Literal	-	Emotionality
Bizzoni[10]	2018	MW	200 Metaphors	Wordnet	Paraphrases
Tweets[106]	2019	MW	2500 Phrases	Emotional Brexit	-

Table 2: Metaphor datasets ordered chronologically.

3.1.2 Metaphor Detection

In this section we discuss metaphor detection systems. The task of metaphor detection can be formulated as a classification or a sequence labeling task. The former involves the construction of a binary classification model that given an input phrase it classifies is as metaphorical or not. The latter formulation requires the construction of a model that given a sentence or phrase it outputs a sequence of labels, one per token/word in the sentence, that represents whether that token is used metaphorically or not.

In this section we will provide a general overview of metaphor detection systems and then will focus more extensively on *state-of-the-art* metaphor detection methods that do not take extended contextual information into account, since textual entailment datasets do not provide context for the examples they contain[16].

We introduce metaphor detection systems categorized by their dominant methodology. Many of the systems discussed in one section may combine more than one of these methodologies. The choice of including one approach into a particular methodology subsection is mostly based on what the authors highlight to be the main contribution of their method. When no such clear preference is evident the decision is arbitrary.

[16] Note, however, that methods that consider only the near context (few sentences) of the metaphor are included in the discussion since MultiNLI[101] may involve more than one sentences in the premise

3.1.2.1 Hyponymy Based One of the earliest approaches on metaphor detection is identifying lack of hyponymy in sentences of the form X is Y[49, 67, 88]. This approach works best for lexical metaphors (nominal, SVO, Adjective Noun) and does not require the expensive training of more modern machine learning based systems. The method proposed in [49] relies on the pre-existing homonymy relations from WordNet[26] and is further improved by [67] through the integration of conceptual mappings through the use of ConceptNet[85].

3.1.2.2 Similarity Based Recall the TroFi[8] dataset presented in the previous subsection. The initial version of weak supervision for the disambiguation of metaphorical or literal use of verbs relies on metrics of semantic and cosine distance. This approach has pervaded into more recent approaches such as [74], where the authors propose an unsupervised fuzzy-rough rule-based classifier heavily reliant on cosine similarity. This work was extended with a clustering algorithm in [75]. Similarity-based metrics have even been employed in state-of-the-art deep learning systems [58] where generated embeddings are compared to web-crawled large scale embedding vectors under the assumption that a metaphoric use of a word will have a larger distance to a literal one, under the assumption that terms are mostly used in their literal sense.

3.1.2.3 Abstractness Based Abstract words are those that are not immediately related to objects that exist to the real world and are in a sense distanced from immediate perception, as are for example the notions of love, intelligence, ideal, etc. On the contrary, concrete words are those that can be immediately observed such as table, dog, and red. In [92] the authors employ a metric of abstractness to classify phrases into metaphoric or literal. The study supports the hypothesis that metaphorical word usage is correlated with the degree of abstractness of the word's context. The abstractness of a word is evaluated by comparing the word to a set of words that are abstract and a set that are concrete projected in a latent space. Then the metrics of abstractness averaged across parts of speech in the dataset are used as features in a classifier. A similar approach was later employed in [29] which augments features of abstractness with imageability and sentiment scores.

3.1.2.4 Word Embedding Based Embedding based approaches have become the most common approach to metaphor detection especially with the advent of deep learning. Shutova et. al. [83] employs visual and textual embeddings to improve metaphor detection systems with visual understandings. This helps identify metaphors like 24 that require visual context.

$$\text{She felt like a black sheep when she met her family} \tag{24}$$

Current state of the art models employ large contextual embeddings in a deep learning pipeline to determine word metaphoricity. One of the first such models[102] combined a Convolutional Neural Network (CNN) with a Bidirectional Long-Short-Term-Memory (BiLSTM) sequential model to perform word-level metaphor classification. The authors used Word2Vec, Part of Speech Tags, and word cluster features as input to their model. The approach was further improved by [28] where contextualized word representations were used as inputs (specifically, Elmo[70] and GloVe[68])).

The current state of the art model in metaphor detection[58] builds of the aforementioned work. We discuss the paper in more detail because it is employed in a later section to evaluate the extent to which large scale entailment datasets contain metaphors. Mao et al. recently proposed the first end-to-end deep learning system that takes into account linguistic theories of metaphors, instead of simply treating the task of metaphor detection as a sequence tagging task analogous to those of Part of Speech Tagging and Named Entity Recognition. Recall from section 2.2.1 on linguistic theories of metaphor the Selectional Preference Violation principle and the Metaphor Identification Procedure. Briefly, Selectional Preference Violation postulates that a metaphor is identified by considering the semantic contrast between a target word and its context. The Metaphor Identification Procedure, on the other hand, takes into account the discrepancy of the literal meaning of the word and the meaning it takes within the sentence.

We first consider the model that is based on Metaphor Identification Procedure. The authors use a BiLSTM to model contextual meaning as a hidden state. More precisely, the contextual meaning of a word is encoded by its backward and forward contexts, as well as the word itself. Then, this contextual meaning is compared to the literal meaning which is encoded by a pretrained embedding, and in particular the GloVe embeddings[68]. Note that GloVe embeddings are trained on web-crawled data around the Web. This approach relies on the hypothesis that a word mostly appears with its literal meaning.

The authors also propose a Selection Preference Violation Model, which in fact outperforms the previous model on most datasets. In this model, the word representation is compared to its context. The target word representation is again a BiLSTM hidden state. However, this time the context is composed with an attention based mechanism, that takes into account both the n-left and n-right words from the target word. The left and right context are computed using a multi-head attention mechanism. In this way, the authors avoid problems caused by context that is further away from the metaphor itself and may be essentially noise for the model.

The following table includes information on the performance of the model. In section 4.2 we employ this model on MultiNLI[101] and empirically observe a significant drop in performance. This indicates that the datasets used for metaphor detection should be improved to better match the distribution of metaphors in the wild (pun intended).

Model	Precision	Recall	F1	Accuracy
MIP-Based	71.8	76.3	74.0	93.6
SPV-Based	73.0	75.7	74.3	93.8

Table 3: State of the Art Metaphor Detection Performance

3.1.3 Metaphor Interpretation

The task of metaphor interpretation can be formulated in different ways that relate closely to theories of metaphor discussed in section 2.2.1; these include generating a literal paraphrase (substitution theory), identifying the underlying conceptual mapping, and identifying the properties transferred (class inclusion theory).

3.1.3.1 Paraphrase Based One of the first computational models for the generation of paraphrases focused on verb substitution[82]. The author suggests a model that ranks potential verb substitutes extracted from WordNet[26] and verifies whether their use is literal through a filter inspired from selectional preference theory. In a later paper[12] this approach is extended with a final filter that ensures that paraphrases maintain the original meaning. An embedding based model that does not rely on pre-existing databases[84] was later proposed. In this embedding-based approach the candidate paraphrases are generated though a probabilistic model that generates literal verb candidates that are again filtered using a selectional preference-based approach. Most relevant to this study, [63] devise a metaphor paraphrase method by retraining vector based textual entailment models on a dataset of 327 metaphors.

3.1.3.2 Conceptual Mapping Based One of the first models that selects the underlying conceptual mapping given a metaphoric expression was proposed in [66] where the author suggests a Bayesian network powered by a pre-existing concept mapping database to probabilistically rank conceptual mappings given the input, through a source-target domain mapping. Later work[2], employs an ontology-based database and a frequency analysis to identify underlying conceptual mappings in English and Chinese. The work was further refined[3] to use WordNet data to improve accuracy on low frequency mappings. Similarly, [59] infer the underlying conceptual mapping of a metaphor through the identification of systematic selectional-preferences through the use of Internet scraping. More recent work [79] employs a simple deep neural network to map input sentences to one of 77 candidate conceptual mappings.

3.1.3.3 Class Inclusion Based In an attempt to support the psycholinguistics hypothesis that conventional metaphors are interpreted in the same way as literal statements, Walter Kintsch proposed a latent semantic analysis-based model [46] to extract the features and properties transferred from one class to the other in metaphoric (and literal) statements. Similarly, in [103] the authors propose a model that extracts potential properties that are transferred to the target domain in nominal metaphors through a word association procedure that examines the co-occurrences of concepts in a latent space. Akira Utsumi and Maki Sakamoto investigate the application of latent space analysis for SVO Japanese metaphors[94] through the employment of a manually constructed verb corpus in which the associated verbs are augmented with an online thesaurus. More recently, [76] extended Kintsch's work through the employment of affective features for the identification of transfer features of nominal metaphors by using pretrained Word2Vec embeddings to extract the latent emotions in the source domain.

3.2 Textual Entailment

3.2.1 Datasets

Textual entailment is a task that has gotten significant attention by the computational linguistics community and as such a series of datasets were developed over the years.

In 2005, the network of excellence in Pattern Analysis, Statistical Modelling and ComputationAl Learning (PASCAL)[17] published the first Recognizing Textual Entailment (RTE) dataset[19].

The same network (PASCAL) developed two more such datasets after the first was met with significant attention. Those datasets were entitled RTE2, RTE3, respectively[32]. In fact RTE3 was translated in German and Italian.

In 2008, the National Institute of Standards and Technology (NIST)[18] developed the fourth RTE dataset[15]. Following the initial dataset NIST published three more RTE datasets RTE5, RTE6,and RTE7 respectively [6, 7].

With the advent of deep learning and the use of large corpora, a new textual entailment dataset was required. In 2015, the Stanford Natural Language Inference (SNLI) corpus[13] was published which includes 570k manually annotated English textual entailment examples.

MultiGenre NLI (MultiNLI) corpus[101] was published containing 433k examples of annotated pairs. It was built to match the SNLI corpus style and potentially be used concurrently.

Finally, the Cross-Lingual NLI Corpus (XNLI)[18] includes a set of 5,000 test and 2,500 development pairs extracted from the MultiNLI corpus and translated into 14 languages: French, Spanish, German, Greek, Bulgarian, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, Hindi, Swahili, and Urdu.

There exist other datasets that are tangential to textual entailment and are more specific towards its particular applications. One such example is the Question Answering NLI Dataset (QNLI)[77,

[17] http://www.pascal-network.org/
[18] https://www.nist.gov/

96] which consists of Question-Answer pairs, and information regarding whether the given answer indeed contains the ground-truth answer to the question. Another such dataset was published in the context of the 8th Recognizing Texual Entailment Challenge at SemEval 2013. This dataset is called the Joint Student Response Analysis dataset which contains a series of questions with multiple answers - one reference answer and one student answer - with information as to whether the student answer entails the reference answer. This dataset is particularly interesting because it is application-specific and aims to broaden education capabilities.

3.2.2 Deep Learning Textual Entailment Approaches

In this section we discuss the best performing deep learning architectures for TE. We focus on the models that are used in section 5.2. The models are discussed in chronological order (by publication date on Arxiv).

BERT[23] BERT stands for **B**idirectional **E**ncoder **R**epresentations from **T**ransformers and was introduced by Google AI in 2018. It provides an unsupervised way to pretrain a deep learning representation of text that can then be finetuned with minimal architectural changes for particular tasks. In fact, at the time of publishing, BERT achieved state of the art performance in 11 tasks - one of which was textual entailment. The input to the model is tokenized and represented as a sequence of token level embedings. Each token level embedding comprises of three separate embeddings: a token embedding, a segment embedding, and a positional embedding, which capture information about the given token, its role as a premise or hypothesis, and position, respectively. The input is then passed on a multilayer transfomer encoder introduced in [95] to generate the output features that are subsequently passed into the classification layers.

BERT was the first model to achieve true bidirectionality in training, primarily through the optimization of a Masked Language Model (MLM) task. Bidirectionality is the property of employing information from both the left and right side context of the token being processed. MLM is defined as follows: given a sentence, mask a subset of its tokens, and create a model that can predict the masked component. Apart from MLM, BERT optimizes another sentence-level objective simultaneously called Next Sentence Prediction (NSP). This task requires the model to predict what will be the next sentence from a set of provided options, given the current sentence. NSP aims to improve performance for sentence level tasks, such as TE. Nevertheless, NSP was shown to be a weak optimization metric and was omitted by later architectures[57, 104].

XLNet[104] XLNet was created in collaboration between Google AI and Carnegie Mellon University with the aim to address some of the shortcomings of BERT. By doing so, XLNet outperformed BERT in 20 natural language tasks, including natural language inference.

We first discuss the shortcomings of BERT that are addressed by XLNet. Recall that BERT masks a random subset of the sentence tokens and optimizes the parameters with respect to predicting the masked tokens. There are two main problems with this approach: Firstly, the masks are encoded with artificial symbols that are absent from natural language thus inducing a discrepancy between the pre-training and fine tuning inputs. In addition, the masked input requires BERT to assume that the tokens of a sentence are independent of each other, which is oversimplified for the higher-order dependencies that are prevalent in natural language.

XLNet addresses these issues through the optimization of the network using an objective inspired from Permutation Language Modeling (PLM). PLM is an autoregressive procedure that achieves bidirectionality by considering all permutations of a sentence. Autoregression is the procedure of predicting the next value in a sequence given the previous values. Let x_t be the t'th token in a sentence; in an autoregressive model we would provide all tokens $x_{i<t}$ in the model and we would require it to predict x_t. This is clearly not bidirectional; in order to include bidirectionality PLM considers all possible permutations of the sentence and as such both the tokens to the left and to the right of x_t eventually inform the optimization. Notice how this procedure does not require the use of any special mask tokens and as such it does not induce a discrepancy between pre-training and fine-tuning inputs. Note than in order for the model to learn useful representations, a position encoding in the sentence is required to be maintained in the token embeddings. Further, note that the absence of masked input tokens eliminates the independence assumption between sentence terms.

Another significant difference between XLNET and BERT is the fact that XLNET is designed to tackle sequence to sequence task which follows naturally from PLM. This is because PLM produces a prediction for each token in the input sequence. BERT also produces a representation for each token in the sequence, but that is later passed in a final classification network.

Despite its improved performance, XLNet is very resource-taxing since it requires significantly more computational resources, time, and data for convergence.

RoBERTa[57] RoBERTa stands for **R**obustly **O**ptimized **BERT** Pretraining **A**pproach. As the name implies, RoBERTa is a vigorously tuned version of BERT. RoBERTa improves upon the BERT model through tuning, training and maintaining the aspects of BERT that lead to highest boost in performance. The authors trained the model for a longer time with larger batches over more data, removed the next sentence prediction objective, trained on longer sequences, and finally dynamically changed the masking pattern applied to the training data. As a result, RoBERTa outperformed BERT and XLNET on all Glue[96] benchmarks, including natural language inference.

ALBERT[54] ALBERT stands for **A L**ite **BERT** and, as its name implies, is a reduced version of BERT. The base version of ALBERT contains 12 million parameters, which is an order of magnitude less than those of BERT. Despite its smaller size ALBERT outperforms BERT and RoBERTa in a series of tasks, including natural language inference. BERT's parameters are reduced in two ways: through factorized embedding parameterization (FEP) and cross-layer parameter sharing (CLPS). FEP involves the decomposition of the large vocabulary embedding matrix into two small matrices, achieving a separation between the size of the hidden layers and the parameters of the vocabulary. CLPS allows the sharing of parameters across layers, thus not requiring the parameter size to increase analogously to the network's depth. Apart from optimizing the size of the model, the authors also introduced a self-supervised optimization task, sentence order prediction (SOP) to replace NSP which was shown to be relatively ineffective. SOP requires the model to predict whether two sentences are in order or not, thus allowing the model to learn sentence coherence[19].

DistilBERT[80] As the name suggests, DistilBERT learns a distilled version of BERT. Distillation[38] is a procedure by which we transfer the behavior of a network to a network with fewer

[19] Note that the authors of ALBERT show that SOP can solve NSP to a reasonable extend, whereas the opposite is not true.

parameters. This is achieved by using two optimization metrics: the loss between the ground truth labels and the predicted labels by the small model, and the loss between the logits[20] of the large and small models. Since DistilBERT mimics BERT's performance, it is not expected to perform better, however, it does maintain a high portion of the original performance.

Bart[56] BART, introduced by Facebook AI, is an autoencoder network. Unlike conventional networks which optimize the loss between predicted output x with ground truth label y, autoencoders aim to reconstruct the input and thus optimize the loss between the reconstructed $\hat{x}$ and input x representations. BART is a denoising autoencoder; given input text x, the input is randomly corrupted to x_c, and then the model learns to reconstruct the original input x from x_c. Note that denoising autoencoders differ from MLM (employed in BERT) with respect to the output: MLM outputs *only* the masked word[21] whereas denoising autoencoders aim to reconstruct the whole input. Like BERT, BART employs a transformer encoder, but instead of being passed on classification layers, the output from the encoder is passed on an autoregressive network similarly to XLNET[104]. The authors experiment with different ways of corrupting the input to the encoder, the most successful of them being text-infilling. Text-infilling involves the masking of randomly sampled segments from the input, with lengths drawn from a Poisson distribution. Finally, the model is optimized with respect to the cross entropy loss between the output and the original input.

Even though BART outperforms BERT in many tasks, especially sequence to sequence tasks, it performs worse on TE benchmarks.

4 Metaphoric Textual Entailment Dataset

4.1 Recast Datasets

In this section we discuss the construction of a metaphoric TE dataset used for fine-tuning and evaluation in section 5.

[20] Logits are the confidence metrics associated with each possible prediction of a network.

[21] It can be thought as a multiclass classification where the vocabulary consists of the classes

4.1.1 Dataset Construction

In this work we use the definition of the term *recasting* as used in the paper [71]. Specifically, the term *recasting* is broadly used to denote the act of *"leveraging existing datasets to create NLI examples"*.

The metaphoric natural language entailment dataset was constructed through recasting two datasets available from earlier work [90][10].

We first discuss the dataset available by [10]. In this paper, the authors construct a corpus of 200 sets of 5 sentences, each containing one reference to a metaphorical sentence and four ranked candidate paraphrases. Their goal is similar to that of this study, with the main difference being the task in question; they focus on graded paraphrase acceptability[22]. In the original dataset, each set of 5 sentences contains a metaphoric sentence and four paraphrase candidates annotated on a scale of 1 to 4, indicating the degree to which a sentence is a paraphrase of the original metaphoric sentence, where 4 stands for exact paraphrase.

The four candidate paraphrase sentences accompanying a metaphoric sentence typically exhibit the following pattern

- Two sentences cannot considered paraphrases in any way, very often due to the fact that they contradict the original sentence.
- Two sentences cannot be paraphrases, but show some degree of semantic similarity to the original sentence.
- Two sentences could be considered paraphrases but have some significant differences in style of content. In other words, they could be viewed as *weak paraphrases*.
- At least one sentence can be considered as a *strong paraphrase*, namely it reiterates the content of the original sentence almost exactly. Such sentences are good candidates to be recast to an entailment pair.

[22] Graded paraphrase acceptability is the task of determining how semantically close one sentence is to another.

It is easy to see how such a dataset can be easily recast to a TE dataset. Essentially, we can consider each sentence and each of the candidate paraphrases as a sentence pair for a TE classification problem. Indeed we manually iterated over all these pairs to construct a dataset of 800 such pairs. From table 4 we observe some variability in the recast dataset. However, it should be noted that entailment pairs significantly dominate the distribution. This is a common consequence of recasting a dataset created for a different task.

Pair Type	Percentage of Dataset	Absolute Number of Pairs
Entailment	45.62%	365
Contradiction	33.75%	270
Neutral	20.63%	165
Total	100%	800

Table 4: Entailment annotations from Bizzoni (2018)[10] recast dataset.

Notice that the candidate paraphrase sentences are supposed to be literal by design, and as such the generated natural language entailment pairs have metaphoric premises and literal hypotheses. However, it is important to note that the separation between metaphoric language and literal language is not clear cut. Often, the same sentence could be considered literal in one context and metaphoric in an other. We explain the relevance of this observation through an example from the original dataset[10].

```
Original Phrase:

    she cut him down with her words


Candidate Paraphrases:

    1. she cheered him up with her words

    2. she left him bored with her words

    3. she put him down with her words

    4. she told him things that made him sad
```

First, notice that the original phrase is metaphoric, since words cannot literally be used to cut anything, let alone a human being. There are a few comments that arise from this particular example, but I first want to guide your attention to the third candidate paraphrase sentence “she put him down with her words”. The authors of the dataset interpret this phrase as literal, however, one can see that it could be considered metaphorical, since words, in the same way that they do not “cut” a person, they do not also literally “put them down”. Now, it is granted that the term “put down” has been pervasively become a phrasal verb to mean “suppress” or “defeat” someone. However, the same argument could be made for the phrase “cut down”. However, for the purposes of this study, we maintain that no matter how common the phrase “cut down” has become, it is still a metaphor, albeit a conventional metaphor.

Apart from annotating the sentence pairs for the entailment task, we further annotate the premise sentences according to metaphor type (or simile). In fact, we find a few (6) premises that were similes and not metaphors. A simile is defined as a figure of speech comparing two unlike things and is introduced by a signaling word, such as ‘like’ or ‘as’. In fact, certain views look at metaphors as an abbreviated simile [61]. At its simplest form, this theory supposes that a simile

Her cheeks are like roses. (25)

can be rewritten as a metaphor

Her cheeks are roses. (26)

without the analogy signal word.

The types of metaphors annotated in the dataset were Subject-Verb-Object (SVO), Adjective-Noun (AN), Nominal (NO), and Multiword (MW) metaphors. Table 5 shows their variability.

Metaphor Type	Percentage	Absolute Number
SVO	37.00%	74
AN	17.50%	35
NO	29.50%	59
Simile	3.00%	6
MW	13.00%	26
Total	100%	200

Table 5: Metaphor Types in Bizzoni Dataset (Premises Only)

The second dataset we recast to entailment pairs was taken from [90]. This dataset consisted of 200 English metaphoric sentences, 200 English literal sentences, 200 Russian metaphoric sentences, and 200 Russian literal sentences. For the purposes of this study we were only concerned with the 200 English metaphoric sentences which were further annotated by whether the metaphor was a Subject-Verb-Object sentence or and Adjective-Noun phrase, and the division was split exactly in half (100 SVO, 100 AN sentences).

Each of the metaphoric sentences in the original datasets were used as premises for the new entailment dataset. This resulted in 199 premises, as one sentence in the original dataset was not interpretable by any of the three native English language speakers asked[23]. The omitted sentence was “If you walk documents into the office you can also pay by VISA or MasterCard”. For each of the 199 premises, we manually created three literal hypotheses with corresponding labels (entailment, neutral, and contradiction). Since those hypotheses were constructed from scratch we ensured that there was an equal (33%) representation for each annotation. Moreover, we were able to limit as much the use of borderline metaphoric language in the hypotheses, something that was not possible for the Bizzoni [10] dataset since the hypotheses were recast from the paraphrase candidates.

4.1.2 Data Statistics

We collected a dataset of 800 pairs from [10] and 597 pairs from [90], in total **1397** pairs, annotated by entailment relationship and metaphor type.

Label	Percentage	Absolute
entailment	40.37%	564
contradiction	33.57%	469
neutral	26.06%	364
total	100%	1397

Table 6: Label variability

Metaphor Type	Percentage	Absolute (pairs)
SVO	42.45%	593
AN	31.15%	440
NO	16.89%	236
Simile	1.72%	24
MW	7.44%	104
Total	99.65%[a]	1397

Table 7: Metaphor Type Premise Variability (counted in terms of total pairs)

[a] Rounding error

Table 8: Dataset Variability

Figure 2 shows the length distribution of the sentences in the metaphoric dataset. We can see that the sentence length is relatively short, supposedly facilitating inference, since deep learning

[23] All three have at least a bachelors degree from an English-speaking University.

models seem to perform better for shorter sentences [65]. For example, the longest sentence in the SNLI dataset contains more than 40 tokens, whereas the longest sentence in our dataset contains just below 30 tokens. In general, we see similar length distributions with popular datasets for the same task such as SNLI [13], and MultiNLI [101].

We further provide a break down of the most frequent words used metaphorically in the premises in table 10. We can observe that there is significant variability, with only one term (the verb ‘feel’) appearing more than 3 times.

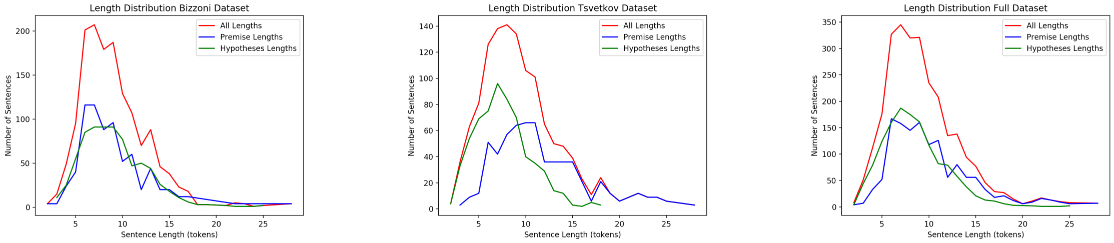


Figure 2: Sentence Length Distribution

Statistic	Full Dataset	Tsvetkov	Bizonni
Premise Maximum Length	28	28	28
Hypotheses Maximum Length	25	25	25
Premise Mean Length	10.12	10.61	9.73
Hypotheses Mean Length	8.31	8.10	8.59
Total Mean Length	9.22	9.36	9.16
Vocabulary Length (uncased)	3315	2387	1446

Table 9: Key Statistics About Metaphor Dataset

4.1.3 Annotation Agreement

Apart from the variability, both in terms of labels as well as metaphor types, the dataset exhibits high quality in terms of annotation agreement. Two more annotators were presented the pairs of sentences randomly shuffled, and were asked to label them as entailment, contradiction, or neutral. Both annotators are undergraduate students in English speaking universities but second language English speakers. The annotators were presented the instructions shown in figure 7 (see Appendix) which were adapted from the instructions provided to the annotators of the MultiNLI dataset[100]. We achieved a significant portion of unanimity, reaching up to 78.95%. A break down of agreement percentage by label and metaphor type can be found in the table 14.

Verbs (SVO)

Verb	Frequency
feel	4
turn	3
break	3
kill	3
shoot	3
plant	3

Nouns (AN)

Noun	Frequency
brain	3
mood	3
humor	2
idea	2
laughter	2
temper	2
mind	2
thinking	2
smile	2
excuse	2
budget	2
rhetoric	2

Adjectives (AN)

Adjective	Frequency
black	3
dark	3
heavy	3
sweet	3
bright	2
foggy	2
red	2
sour	2
stormy	2
warm	2
bitter	2

Table 10: Most Frequent Terms in SVO and AN metaphors (premises only)

Label	Percentage	Absolute
entailment	90.43%	510
contradiction	72.92%	342
neutral	68.96%	251
total	78.95%	1103

Table 11: Annotator Agreement By Label

Metaphor Type	Percentage	Absolute (pairs)
SVO	80.61%	478
AN	80.91%	356
NO	70.34%	166
Simile	91.67%	22
MW	77.88%	81
Total	78.95%	1103

Table 12: Annotator Agreement By Metaphor Type

Source Data	Percentage	Absolute (pairs)
Tsvetkov	92.62%	503
Bizonni	81.25%	600
Total	78.95%	1103

Table 13: Annotator Agreement By Source Data

Table 14: Annotator Agreement Break Down

4.2 Metaphor Extraction from Large Datasets

Before deciding to recast existing datasets to construct the metaphoric entailment dataset we experimented with extracting metaphors from larger datasets. Specifically, we were concerned with MultiNli, since SNLI premises were constructed from image captions, making it unlikely for a significant amount of metaphoric sentences to be included in the set.

Unfortunately, state of the art metaphor detection models did not yield an accurate enough output to be useful. The results are nevertheless reported, partly to motivate the idea that metaphor detection is an area that needs to be further investigated. This is especially true, given the fact that the computational linguistics community seems to have taken a path of developing separate models for metaphoric and literal language.

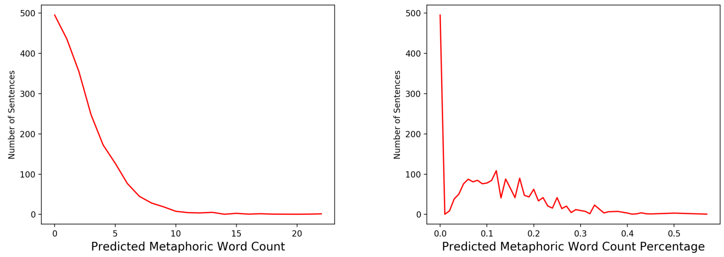


Figure 3: Information about the distribution of predicted number of metaphoric words in MultiNLI dataset. The left graph shows the distribution of the absolute number of metaphoric words predicted per sentence whereas the right one show the distribution of the percentage of predicted metaphoric words in terms to the total words per sentence.

The test set of MultiNLI contains a total of 2082 distinct premises. We used a state of the art metaphor detection system[58] trained on the VUA Metaphor Dataset[55]. Given a sentence and its part-of-speech tag-sequence the model outputs whether each token in the sentence is used metaphorically or not. The part-of-speech tag-sequences were generated using the Allenlnp constituency parser which implements this [43] model. The left graph in figure 3 shows the predicted number of metaphoric words in the MultiNLI dataset. We can see that a large portion of the sentences (24.36%) are predicted to have no metaphoric use of words. Due to the variability of lengths of the sentences we also include a distribution of the percentage of metaphoric words per sentences predicted by the model in the right graph of 3.

We observe that the model incorrectly categorizes many words as metaphoric (high rate of false positives). For example, the following sentence *”But the state does arguably have an interest compatible with the First Amendment in stipulating the way those media are used and Fiss discussion of those issues is the least aggravating in his book.”* as containing 8 metaphorically used words. This example is one of many in the dataset that were mistakenly classified.

Even though the model seems to overestimate the number of metaphors in the dataset, and thus makes it difficult to generate a reliable metaphoric dataset using this procedure, this process allows us to evaluate the extent to which metaphoric use of language exists in the MultiNLI dataset. This observation makes the task of evaluating entailment models’ performance on metaphoric sentences more interesting. A small subset of such sentences from the test set are listed in table 15.

Metaphors in MultiNLI
That is as the discount is increased in steps the cost to the Postal Service of sorting the mail that becomes workshared on step 4 is probably greater than the cost of sorting the mail that becomes workshared on step 3 . *This assumption will be relaxed* in Part III below where larger discount changes are considered .
The Times says this tracking list is drawn up from information from bookstores but publishers say they routinely call up the Times to tip them off to books *selling with increasing momentum* so that they can be added to the tracking list .
will never be doused Brit Hume Fox News Sunday; Tony Blankley Late Edition;Robert Novak Capital Gang; Tucker Carlson The McLaughlin Group. The middle way is best expressed by Howard Kurtz NBC is Meet the Press he scolds Brill for undisclosed campaign contributions and for *overstretching his legal case* against Kenneth Starr but applauds him for *casting light on the media*.
You *claw your way into* a position to get your calls returned by actually breaking stories but that reward is empty.
Critics praise Goodman’s finely honed descriptive abilities and instinctive grasp of familial dynamics the ways in which *dreams and emotional habits are handed down* ...
I can’t help but wonder if Shuger thought to ask himself a few simple questions before *launching his attack questions* such as did Tripp ask to be moved to her current job?
Pro Microsoft analysts spin this as a heroic sacrifice removing the lightning rod whose seemingly disingenuous testimony has ostensibly *driven the DOJ to the verge* of demanding the company is breakup .

Table 15: Metaphors in MultiNLI Development Split. The metaphoric phrase is enclosed between asterisks ’*’

5 Experiments

We evaluate a series of deep learning natural language entailment models on the metaphoric dataset constructed in section 4. We perform the following experiments on a series of deep learning entailment models presented in section 3.2.2. All fine-tuning consists of at least 3 epochs with the default parameters provided by the authors of the models.

1. Fine-tune on MultiNLI and evaluate on MultiNLI
2. Fine-tune on MultiNLI and evaluate on SNLI
3. Fine-tune on MultiNLI and evaluate on Metaphoric Dataset
4. Fine-tune on MultiNLI and Metaphoric Dataset[24] and evaluate on Metaphoric Dataset
5. Fine-tune on MultiNLI and Metaphoric Dataset and evaluate on MultiNLI
6. Fine-tune on MultiNLI and Metaphoric Dataset and evaluate on SNLI
7. Fine-tune on MultiNLI and Metaphoric DatasetI and evaluate on Metaphoric Dataset
8. Fine-tune on MultiNLI and Metaphoric Dataset and evaluate on SNLI

[24] Note that there is not enough data to only fine tune on metaphors, so we use metaphors with MultiNLI.

9. Fine-tune on MultiNLI and Metaphoric Dataset and evaluate on MultiNLI

We also consider a baseline of hypotheses-only classification which has been suggested to outperform majority class baselines[72]. The study further suggests that datasets may exhibit statistical irregularities that may allow models to achieve performance that is higher than what should be expected without the premises. We show that our dataset seems to not exhibit such irregularities since the accuracy on hypotheses-only experiments is quite low compared to that of SNLI, and MultiNLI datasets.

1. Fine-tune on MultiNLI Hypothesis Only and evaluate on Metaphor Dataset Hypotheses Only.
2. Fine-tune on MultiNLI and Metaphor Dataset Hypothesis Only and evaluate on Metaphor Dataset Hypotheses Only.

5.1 Hypotheses Only Baselines

We consider the hypotheses-only baseline[72] to ensure that our dataset does not exhibit any hypothesis level irregularities that allow the model to achieve high classification accuracy without considering the premises. Table 16 reports the accuracy achieved by the baseline. We consider two hypotheses-only baselines: first we only train on the MultiNLI dataset and evaluate on the entire metaphoric dataset, and second we train on both MultiNLI and the train split of the metaphoric dataset and evaluate on the test split of the metaphoric dataset. We perform each experiment for the second baseline (MultiNLI and Metaphoric Dataset training) three times and report the average. Each time we randomly select a new split of the train and evaluation set. We further examine the accuracy achieved for each of the metaphor types. For the metaphor type specific baselines we train on the entire metaphoric train split but only consider specific types in evaluation. The models were allowed to run for up to 20 epochs, or end early as follows: if the test accuracy on falls the learning rate is decreased by a factor of two, and the training stops when the learning rate is less than 0.00001.

Finetune DS	Accuracy
MultiNLI (all)	50.46%
MultiNLI+MetNLI (all)	52.82%
MultiNLI (NO)	47.85%
MultiNLI+MetNLI (NO)	49.54%
MultiNLI (SVO)	51.36%
MultiNLI+MetNLI (SVO)	50.48%
MultiNLI (AN)	52.00%
MultiNLI+MetNLI (AN)	52.80%
MultiNLI (MW)	42.43%
MultiNLI+MetNLI (MW)	45.20%

Table 16: The MultiNLI-trained baseline was computed by evaluating on the whole metaphoric dataset. The MultiNLI+MetNLI baseline reported, is the average of the results obtained by training on MultiNLI and a train split from the metaphoric dataset, and evaluating on the test metaphoric dataset. Each computation was performed in a new split, randomly generated (3 separate splits). When considering separate types of metaphors, the only alteration is in the testing procedure to only include metaphors of this type from the test set (or whole set in the MultiNLI only training baseline).

5.2 Fine-Tuning Experiments

In this section we report results on fine-tuning state of the art natural language entailment systems on the metaphoric dataset.

5.2.1 Quantitative Results

We first finetune the models on the train split of MultiNLI and evaluate them on the development test of MultiNLI[25] and the test set of SNLI. We then proceed to fine tune on MultiNLI and the metaphoric dataset. In order to produce more reliable results, we consider three random train/test splits of the metaphoric dataset and report the average and the sample standard deviation. Notice that we maintain an 80/20 ratio within metaphor types as well, so that we can ensure that the test sets have consistently the same amount of each type of metaphor. During finetuning we cap the maximum epochs to five and only stop fine-tuning early if an epoch has been completed and there are two consecutive drops in evaluation accuracy. Tables 17 and 18 present the results of the experiments. The first column identifies the model used, the second the fine-tuning dataset(s), the third the evaluation dataset, and the rest indicate the accuracy. Table 19 includes a break down of the results by metaphor type[26].

[25] The test set labels are not publicly available.

[26] Breakdown for simile is not provided since the sample was too small to show meaningful results.

The empirical results indicate that the models perform significantly worse on the metaphoric dataset when trained on MultiNLI compared to their performance on both MultiNLI (the train distribution) and SNLI.

All models decrease in performance when evaluated on SNLI after being trained on MultiNLI, with BART achieving the smallest gap. The model with best generalization on metaphors without finetuning on the metaphoric dataset is Roberta with accuracy about 64%. After finetuning on the metaphoric dataset, however, the best performing model is BART, followed by XLNET, achieving almost 80% accuracy. This may indicate that autoregressive procedures are more suited for transfer learning. In fact, the transferring ability of autoregressive procedures has been investigated in the literature[24, 17].

As expected DistilBert and ALBERT are the worst performing models on the metaphoric data. This is not surprising since those models are lighter versions of BERT and it would be odd for them to adapt on a new dataset better than the original model.

Finally, we observe that all models drop in performance on both MultiNLI and SNLI after being tuned on metaphoric data. We consistently observe a better performance on MultiNLI over SNLI since SNLI is from a completely different distribution. Also since SNLI premises are constructed from image captions, it is unlikely to benefit from the metaphoric examples in training.

The data offer a break down in performance by metaphor type. The most obvious conclusion that can be drawn from table 19 is the fact that the models seem to consistently perform best on adjective-noun metaphors. One reason for this may be that the hypotheses associated with adjective-noun sentences can be resolved through information of the rest of the sentence, despite an effort to address the metaphoric aspect of premises on hypothesis creation. Moreover, we observe consistent improvement in all metaphor types when fine-tuning on the metaphoric dataset.

5.2.2 Qualitative Results

In this section we consider different examples that were classified correctly or incorrectly by all models.

The models were able to consistently categorize some metaphors even before fine-tuning on

Model	Finetuning	Eval	Acc	σ
Bert	MNLI	Met	0.61	0.02
Bert	MNLI	MNLI	0.847	-
Bert	MNLI	MNLI mm	0.845	-
Bert	MNLI	SNLI	0.794	-
Bert	MNLI+Met	Met	0.755	0.012
Bert	MNLI+Met	MNLI	0.782	0.005
Bert	MNLI+Met	MNLI mm	0.775	0.007
Bert	MNLI+Met	SNLI	0.635	0.026
DistilBert	MNLI	Met	0.565	0.019
DistilBert	MNLI	MNLI	0.809	-
DistilBert	MNLI	MNLI mm	0.818	-
DistilBert	MNLI	SNLI	0.742	-
DistilBert	MNLI+Met	Met	0.728	0.028
DistilBert	MNLI+Met	MNLI	0.748	0.002
DistilBert	MNLI+Met	MNLI mm	0.741	0.003
DistilBert	MNLI+Met	SNLI	0.575	0.02
Albert	MNLI	Met	0.595	0.019
Albert	MNLI	MNLI	0.849	-
Albert	MNLI	MNLI mm	0.856	-
Albert	MNLI	SNLI	0.804	-
Albert	MNLI+Met	Met	0.711	0.027
Albert	MNLI+Met	MNLI	0.718	0.016
Albert	MNLI+Met	MNLI mm	0.717	0.018
Albert	MNLI+Met	SNLI	0.559	0.096

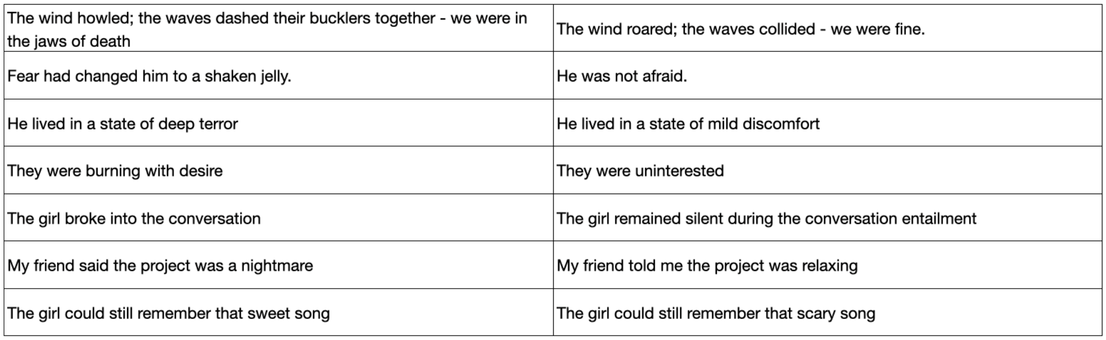
Table 17: Entailment Model Evaluation on Metaphors (Part A)

Model	Finetuning	Eval	Acc	σ
BART	MNLI	Met	0.635	0.022
BART	MNLI	MNLI	0.89	-
BART	MNLI	MNLI mm	0.89	-
BART	MNLI	SNLI	0.875	-
BART	MNLI+Met	Met	0.798	0.038
BART	MNLI+Met	MNLI	0.855	0.004
BART	MNLI+Met	MNLI mm	0.847	0.004
BART	MNLI+Met	SNLI	0.779	0.011
XLNET	MNLI	Met	0.629	0.018
XLNET	MNLI	MNLI	0.871	-
XLNET	MNLI	MNLI mm	0.868	-
XLNET	MNLI	SNLI	0.817	-
XLNET	MNLI+Met	Met	0.784	0.023
XLNET	MNLI+Met	MNLI	0.815	0.013
XLNET	MNLI+Met	MNLI mm	0.8	0.013
XLNET	MNLI+Met	SNLI	0.644	0.023
Roberta	MNLI	Met	0.637	0.017
Roberta	MNLI	MNLI	0.881	-
Roberta	MNLI	MNLI mm	0.879	-
Roberta	MNLI	SNLI	0.845	-
Roberta	MNLI+Met	Met	0.763	0.02
Roberta	MNLI+Met	MNLI	0.815	0.001
Roberta	MNLI+Met	MNLI mm	0.803	0.002
Roberta	MNLI+Met	SNLI	0.662	0.021

Table 18: Entailment Model Evaluation on Metaphors (Part B)

Model	Finetuning	Acc. SVO	Acc. NO	Acc. AN	Acc. MW
Bert	MNLI	0.536 ± 0.042	0.535 ± 0.06	0.759 ± 0.033	0.682 ± 0.045
Bert	MNLI+Met	0.696 ± 0.049	0.819 ± 0.127	0.785 ± 0.042	0.848 ± 0.069
DistilBert	MNLI	0.525 ± 0.024	0.486 ± 0.024	0.645 ± 0.047	0.682 ± 0.045
DistilBert	MNLI+Met	0.709 ± 0.058	0.715 ± 0.073	0.737 ± 0.073	0.833 ± 0.052
Albert	MNLI	0.563 ± 0.02	0.514 ± 0.098	0.693 ± 0.03	0.621 ± 0.026
Albert	MNLI+Met	0.653 ± 0.051	0.757 ± 0.073	0.754 ± 0.055	0.788 ± 0.026
BART	MNLI	0.605 ± 0.024	0.507 ± 0.012	0.724 ± 0.047	0.773 ± 0.079
BART	MNLI+Met	0.752 ± 0.05	0.785 ± 0.043	0.855 ± 0.026	0.894 ± 0.069
XLNET	MNLI	0.517 ± 0.103	0.542 ± 0.055	0.759 ± 0.072	0.652 ± 0.026
XLNET	MNLI+Met	0.739 ± 0.032	0.799 ± 0.084	0.838 ± 0.04	0.818 ± 0.045
Roberta	MNLI	0.592 ± 0.016	0.569 ± 0.087	0.75 ± 0.023	0.652 ± 0.052
Roberta	MNLI+Met	0.704 ± 0.058	0.778 ± 0.098	0.82 ± 0.033	0.864 ± 0.045

Table 19: Performance Analysis by Metaphor Type



The wind howled; the waves dashed their bucklers together - we were in the jaws of death	The wind roared; the waves collided - we were fine.
Fear had changed him to a shaken jelly.	He was not afraid.
He lived in a state of deep terror	He lived in a state of mild discomfort
They were burning with desire	They were uninterested
The girl broke into the conversation	The girl remained silent during the conversation entailment
My friend said the project was a nightmare	My friend told me the project was relaxing
The girl could still remember that sweet song	The girl could still remember that scary song

Figure 4: Premises are on the left and hypotheses on the right.

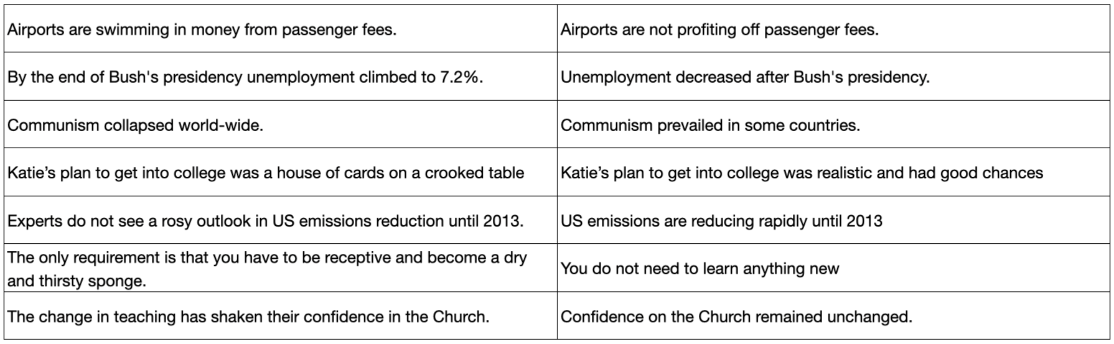
the metaphoric dataset. Table 4 shows some of these examples.

We first consider some sentences that were classified incorrectly by all models but after metaphor fine-tuning were classified correctly by all of them. Table 5 shows some of these examples.

We notice that even after training on metaphors some metaphors that are more abstract are still categorized incorrectly. Table 6 shows some of these examples.

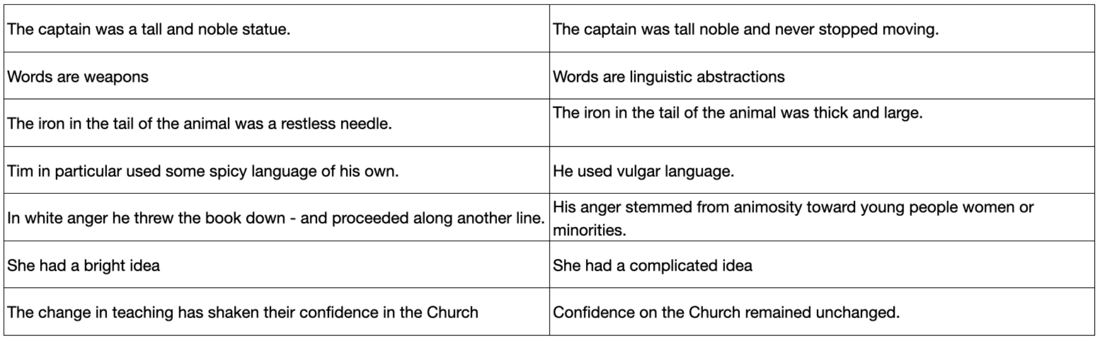
6 Conclusion and Future Work

One of the goals of this study was to evaluate the potential of joint models in terms of processing literal and metaphorical language. We experimented with fine-tuning deep learning entailment systems on a small metaphoric dataset that was constructed specifically for this study, introduced in section 4. The results showed that even with this small dataset the performance of the mod-



Airports are swimming in money from passenger fees.	Airports are not profiting off passenger fees.
By the end of Bush's presidency unemployment climbed to 7.2%.	Unemployment decreased after Bush's presidency.
Communism collapsed world-wide.	Communism prevailed in some countries.
Katie’s plan to get into college was a house of cards on a crooked table	Katie’s plan to get into college was realistic and had good chances
Experts do not see a rosy outlook in US emissions reduction until 2013.	US emissions are reducing rapidly until 2013
The only requirement is that you have to be receptive and become a dry and thirsty sponge.	You do not need to learn anything new
The change in teaching has shaken their confidence in the Church.	Confidence on the Church remained unchanged.

Figure 5: Premises are on the left and hypotheses on the right.



The captain was a tall and noble statue.	The captain was tall noble and never stopped moving.
Words are weapons	Words are linguistic abstractions
The iron in the tail of the animal was a restless needle.	The iron in the tail of the animal was thick and large.
Tim in particular used some spicy language of his own.	He used vulgar language.
In white anger he threw the book down - and proceeded along another line.	His anger stemmed from animosity toward young people women or minorities.
She had a bright idea	She had a complicated idea
The change in teaching has shaken their confidence in the Church	Confidence on the Church remained unchanged.

Figure 6: Premises are on the left and hypotheses on the right.

els on metaphoric dataset rises consistently and significantly. This indicates that TE may be a computationally promising approach to metaphor interpretation.

Furthermore, note that the trade-off between performance in the initial (largely literal) dataset and the metaphoric dataset varies across the models. We believe that this indicates that there is potential for constructing datasets that are purposefully filled with metaphoric language to construct models that can perform well enough in both tasks. Unlike earlier conclusions[63] we find that there is potential of developing systems that can process literal and metaphoric data within the same pipeline. This is particularly beneficial because of the current state of metaphor detection systems, which does not allow them to be reliably applied for the recognition of metaphoric and non-metaphoric uses of language as seen in section 4.2. It is important to note that any such endeavor should distinguish between novel and conventional metaphors, since interpreting the prior is a much harder task even for humans, as studies[51, 50] show. Finally, improving the performance of standard entailment models on metaphors may even help improve their overall performance. Recall that in section 4.2 we showed that large entailment datasets contain metaphoric sentences and as such improving performance on metaphoric sentences could improve overall performance on the task. Another way in which the incorporation of metaphoric interpretations could be beneficial for general-purpose inference models stems from the fact that metaphors provide unique insight in analogical reasoning, which could be helpful for the interpretation of even literal segments as well. As the late Jaime G. Carbonell wrote *"Metaphor is the reflection, on the language medium, of analogical thought processes; as such it provides essential clues of the inner functioning of human inference processes."*[16].

References

[1] Rodrigo Agerri. "Metaphor in textual entailment". In: *Coling 2008: Companion volume: Posters*. 2008, pp. 3–6.

[2] Kathleen Ahrens, Siaw Fong Chung, and Chu-Ren Huang. "Conceptual Metaphors: Ontology-based representation and corpora driven Mapping Principles". In: *Proceedings of the ACL 2003 workshop on Lexicon and figurative language-Volume 14*. Association for Computational Linguistics. 2003, pp. 36–42.

[3] Kathleen Ahrens, Siaw-Fong Chung, and Chu-Ren Huang. "From lexical semantics to conceptual metaphors: Mapping principle verification with wordnet and sumo". In: *Recent Advancement in Chinese Lexical Semantics: Proceedings of 5th Chinese Lexical Semantics Workshop (CLSW-5), Singapore: COLIPS*. 2004, pp. 99–106.

[4] Ion Androutsopoulos and Prodromos Malakasiotis. "A survey of paraphrasing and textual entailment methods". In: *Journal of Artificial Intelligence Research* 38 (2010), pp. 135–187.

[5] Emily Ayoob. "Black & Davidson on metaphor". In: *Macalester Journal of Philosophy* 16.1 (2007), p. 6.

[6] Luisa Bentivogli et al. "The Fifth PASCAL Recognizing Textual Entailment Challenge." In: *TAC*. 2009.

[7] Luisa Bentivogli et al. "The Seventh PASCAL Recognizing Textual Entailment Challenge." In: *TAC*. 2011.

[8] Julia Birke and Anoop Sarkar. "A clustering approach for nearly unsupervised recognition of nonliteral language". In: *11th Conference of the European Chapter of the Association for Computational Linguistics*. 2006.

[9] Julia Birke and Anoop Sarkar. "Active learning for the identification of nonliteral language". In: *Proceedings of the Workshop on Computational Approaches to Figurative Language*. 2007, pp. 21–28.

[10] Yuri Bizzoni and Shalom Lappin. "Predicting human metaphor paraphrase judgments with deep neural networks". In: *Proceedings of the Workshop on Figurative Language Processing*. 2018, pp. 45–55.

[11] Max Black. "XII-METAPHOR". In: *Proceedings of the Aristotelian Society*. Vol. 55. 1. Oxford University Press Oxford, UK. 1955, pp. 273–294.

[12] Danushka Bollegala and Ekaterina Shutova. "Metaphor interpretation using paraphrases extracted from the web". In: *PloS one* 8.9 (2013).

[13] Samuel R. Bowman et al. "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2015.

[14] George R Boys-Stones et al. *Metaphor, allegory, and the classical tradition: ancient thought and modern revisions*. Oxford University Press on Demand, 2003.

[15] Elena Cabrio, Milen Kouylekov, and Bernardo Magnini. “Combining Specialized Entailment Engines for RTE-4.” In: *TAC*. 2008.

[16] Jaime G Carbonell. “Metaphor: an inescapable phenomenon in natural-language comprehension”. In: *Strategies for natural language processing* 415 (1982).

[17] Y. Chung and J. Glass. “Generative Pre-Training for Speech with Autoregressive Predictive Coding”. In: *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2020, pp. 3497–3501.

[18] Alexis Conneau et al. “XNLI: Evaluating Cross-lingual Sentence Representations”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018.

[19] Ido Dagan, Oren Glickman, and Bernardo Magnini. “The PASCAL recognising textual entailment challenge”. In: *Machine Learning Challenges Workshop*. Springer. 2005, pp. 177–190.

[20] Ido Dagan et al. “Recognizing textual entailment: Models and applications”. In: *Synthesis Lectures on Human Language Technologies* 6.4 (2013), pp. 1–220.

[21] Donald Davidson. “What metaphors mean”. In: *Critical inquiry* 5.1 (1978), pp. 31–47.

[22] E Den Boer. “The frequency of original metaphors in literary and nonliterary texts”. In: *biannual conference of the Internationale Gesellschaft für Empirische Literaturwissenschaft, Utrecht University, Utrecht, The Netherlands*. 1998.

[23] Jacob Devlin et al. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[24] Nimit Dhulekar et al. “Seizure prediction by graph mining, transfer learning, and transformation learning”. In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer. 2015, pp. 32–52.

[25] Emily Dickinson. “Hope is the Thing With Feathers-314”. In: *Poetry Foundation* (1924).

[26] Christiane Fellbaum. “WordNet”. In: *The encyclopedia of applied linguistics* (2012).

[27] Joseph L Fleiss. “Measuring nominal scale agreement among many raters.” In: *Psychological bulletin* 76.5 (1971), p. 378.

[28] Ge Gao et al. “Neural metaphor detection in context”. In: *arXiv preprint arXiv:1808.09653* (2018).

[29] Andrew Gargett and John Barnden. “Modeling the interaction between sensory and affective meanings for detecting metaphor”. In: *Proceedings of the Third Workshop on Metaphor in NLP*. 2015, pp. 21–30.

[30] Omar Carlo Gioacchino Gelo and Erhard Mergenthaler. “Unconventional metaphors and emotional-cognitive regulation in a metacognitive interpersonal therapy”. In: *Psychotherapy Research* 22.2 (2012), pp. 159–175.

[31] Dedre Gentner. “Structure-mapping: A theoretical framework for analogy”. In: *Cognitive science* 7.2 (1983), pp. 155–170.

[32] Danilo Giampiccolo et al. “The third pascal recognizing textual entailment challenge”. In: *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*. Association for Computational Linguistics. 2007, pp. 1–9.

[33] Sam Glucksberg. “The psycholinguistics of metaphor”. In: *Trends in cognitive sciences* 7.2 (2003), pp. 92–96.

[34] Sam Glucksberg, Matthew S McGlone, and Deanna Manfredi. “Property attribution in metaphor comprehension”. In: *Journal of memory and language* 36.1 (1997), pp. 50–67.

[35] Pragglejaz Group. “MIP: A method for identifying metaphorically used words in discourse”. In: *Metaphor and Symbol* 22.1 (2007), pp. 1–39.

[36] Hessel Haagsma and Johannes Bjerva. “Detecting novel metaphor using selectional preference information”. In: *Proceedings of the Fourth Workshop on Metaphor in NLP*. 2016, pp. 10–17.

[37] Stephen Halliwell et al. *Aristotle’s poetics*. University of Chicago Press, 1998.

[38] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015).

[39] Dirk Hovy et al. “Identifying metaphorical word use with tree kernels”. In: *Proceedings of the First Workshop on Metaphor in NLP*. 2013, pp. 52–57.

[40] Bipin Indurkhya. *Metaphor and cognition: An interactionist approach*. Vol. 13. Springer Science & Business Media, 2013.

[41] Albrecht Werner Inhoff, Susan D Lima, and Patrick J Carroll. “Contextual effects on metaphor comprehension in reading”. In: *Memory & Cognition* 12.6 (1984), pp. 558–567.

[42] Anthony Jannotta. “Plato’s Theory of Forms: Analogy and Metaphor in Plato’s Republic”. In: *Undergraduate Review* 6.1 (2010), pp. 154–157.

[43] Vidur Joshi, Matthew Peters, and Mark Hopkins. “Extending a parser to distant domains using a few dozen partially annotated examples”. In: *arXiv preprint arXiv:1805.06556* (2018).

[44] Boaz Keysar. “On the functional equivalence of literal and metaphorical interpretations in discourse”. In: *Journal of memory and language* 28.4 (1989), pp. 375–385.

[45] Boaz Keysar et al. “Conventional language: How metaphorical is it?” In: *Journal of Memory and Language* 43.4 (2000), pp. 576–593.

[46] Walter Kintsch. “Metaphor comprehension: A computational theory”. In: *Psychonomic bulletin & review* 7.2 (2000), pp. 257–266.

[47] Zoltan Kovecses. *Metaphor: A practical introduction*. Oxford University Press, 2010.

[48] Zoltán Kövecses. “The scope of metaphor”. In: *Metaphor and metonymy at the crossroads: A cognitive perspective* (2000), pp. 79–92.

[49] Saisuresh Krishnakumaran and Xiaojin Zhu. “Hunting Elusive Metaphors Using Lexical Resources.” In: *Proceedings of the Workshop on Computational approaches to Figurative Language*. 2007, pp. 13–20.

[50] Vicky Tzuyin Lai and Tim Curran. “ERP evidence for conceptual mappings and comparison processes during the comprehension of conventional and novel metaphors”. In: *Brain and Language* 127.3 (2013), pp. 484–496.

[51] Vicky Tzuyin Lai, Tim Curran, and Lise Menn. “Comprehending conventional and novel metaphors: An ERP study”. In: *Brain research* 1284 (2009), pp. 145–155.

[52] George Lakoff, Jane Espenson, and Adele Goldberg. “Master Metaphor List”. In: (1989).

[53] George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.

[54] Zhenzhong Lan et al. “Albert: A lite bert for self-supervised learning of language representations”. In: *arXiv preprint arXiv:1909.11942* (2019).

[55] Chee Wee Leong, Beata Beigman Klebanov, and Ekaterina Shutova. “A report on the 2018 VUA metaphor detection shared task”. In: *Proceedings of the Workshop on Figurative Language Processing*. 2018, pp. 56–66.

[56] Mike Lewis et al. “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension”. In: *arXiv preprint arXiv:1910.13461* (2019).

[57] Yinhan Liu et al. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* (2019).

[58] Rui Mao, Chenghua Lin, and Frank Guerin. “End-to-End Sequential Metaphor Identification Inspired by Linguistic Theories”. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 3888–3898.

[59] Zachary J Mason. “CorMet: a computational, corpus-based conventional metaphor extraction system”. In: *Computational linguistics* 30.1 (2004), pp. 23–44.

[60] Diana McCarthy and John Carroll. “Disambiguating nouns, verbs, and adjectives using automatically acquired selectional preferences”. In: *Computational Linguistics* 29.4 (2003), pp. 639–654.

[61] George A Miller. “Images and models, similes and metaphors”. In: *Metaphor and thought* 2 (1979), pp. 2–25.

[62] Saif Mohammad, Ekaterina Shutova, and Peter Turney. “Metaphor as a medium for emotion: An empirical study”. In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 2016, pp. 23–33.

[63] Michael Mohler, Marc Tomlinson, and David Bracewell. “Applying textual entailment to the interpretation of metaphor”. In: *2013 IEEE Seventh International Conference on Semantic Computing*. IEEE. 2013, pp. 118–125.

[64] Gregory L Murphy. “On metaphoric representation”. In: *Cognition* 60.2 (1996), pp. 173–204.

[65] Shashi Narayan et al. “Split and Rephrase”. In: *EMNLP 2017: Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 617–627.

[66] Srini Narayanan. “Knowledge-based action representations for metaphor and aspect (KARMA)”. PhD thesis. Ph. D. thesis, University of California at Berkeley, 1997.

[67] Yair Neuman et al. “Metaphor identification in large texts corpora”. In: *PloS one* 8.4 (2013).

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

[69] Walker Percy. “Metaphor as mistake”. In: *The Sewanee Review* 66.1 (1958), pp. 79–99.

[70] Matthew E. Peters et al. “Deep contextualized word representations”. In: *Proc. of NAACL*. 2018.

[71] Adam Poliak et al. “Collecting diverse natural language inference problems for sentence representation evaluation”. In: *arXiv preprint arXiv:1804.08207* (2018).

[72] Adam Poliak et al. “Hypothesis Only Baselines for Natural Language Inference”. In: *The Seventh Joint Conference on Lexical and Computational Semantics (*SEM)*. 2018.

[73] Sunny Rai and Shampa Chakraverty. “A Survey on Computational Metaphor Processing”. In: *ACM Computing Surveys (CSUR)* 53.2 (2020), pp. 1–37.

[74] Sunny Rai and Shampa Chakraverty. “Metaphor detection using fuzzy rough sets”. In: *International Joint Conference on Rough Sets*. Springer. 2017, pp. 271–279.

[75] Sunny Rai et al. “Soft metaphor detection using fuzzy c-means”. In: *International Conference on Mining Intelligence and Knowledge Exploration*. Springer. 2017, pp. 402–411.

[76] Sunny Rai et al. “Understanding Metaphors Using Emotions”. In: *New Generation Computing* 37.1 (2019), pp. 5–27.

[77] Pranav Rajpurkar et al. “Squad: 100,000+ questions for machine comprehension of text”. In: *arXiv preprint arXiv:1606.05250* (2016).

[78] Ivor Armstrong Richards. “The philosophy of rhetoric”. In: (1970).

[79] Zachary Rosen. “Computationally Constructed Concepts: A Machine Learning Approach to Metaphor Interpretation Using Usage-Based Construction Grammatical Cues”. In: *Proceedings of the Workshop on Figurative Language Processing*. 2018, pp. 102–109.

[80] Victor Sanh et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”. In: *arXiv preprint arXiv:1910.01108* (2019).

[81] Elena Semino, John Heywood, and Mick Short. “Methodological problems in the analysis of metaphors in a corpus of conversations about cancer”. In: *Journal of pragmatics* 36.7 (2004), pp. 1271–1294.

[82] Ekaterina Shutova. “Automatic metaphor interpretation as a paraphrasing task”. In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2010, pp. 1029–1037.

[83] Ekaterina Shutova, Douwe Kiela, and Jean Maillard. “Black holes and white rabbits: Metaphor identification with visual features”. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016, pp. 160–170.

[84] Ekaterina Shutova, Tim Van de Cruys, and Anna Korhonen. “Unsupervised metaphor paraphrasing using a vector space model”. In: *Proceedings of COLING 2012: Posters*. 2012, pp. 1121–1130.

[85] Robert Speer and Catherine Havasi. “Representing General Relational Knowledge in ConceptNet 5.” In: *LREC*. 2012, pp. 3679–3686.

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

[87] Gerard Steen. “Towards a procedure for metaphor identification”. In: *Language and literature* 11.1 (2002), pp. 17–33.

[88] Chang Su, Shuman Huang, and Yijiang Chen. “Automatic detection and interpretation of nominal metaphor based on the theory of meaning”. In: *Neurocomputing* 219 (2017), pp. 300–311.

[89] Paul H Thibodeau and Lera Boroditsky. “Measuring effects of metaphor in a dynamic opinion landscape”. In: *PloS one* 10.7 (2015).

[90] Yulia Tsvetkov et al. “Metaphor detection with cross-lingual model transfer”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014, pp. 248–258.

[91] Mark Turner and George Lakoff. “More than cool reason: A field guide to poetic metaphor”. In: *Journal of Women s Health* (1989).

[92] Peter D Turney et al. “Literal and metaphorical sense identification through concrete and abstract context”. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2011, pp. 680–690.

[93] Justina Urbonaitė et al. “Metaphor identification procedure MIPVU: an attempt to apply it to Lithuanian”. In: *Taikomoji kalbotyra* 7 (2015), pp. 1–25.

[94] Akira Utsumi and Maki Sakamoto. “Indirect categorization as a process of predicative metaphor comprehension”. In: *Metaphor and Symbol* 26.4 (2011), pp. 299–313.

[95] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.

[96] Alex Wang et al. “GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding”. In: In the Proceedings of ICLR. 2019.

[97] WANG Wen-bin. “On the Subjectivity and Subjective Self-negotiation in Metaphor Interpretation [J]”. In: *Journal of Foreign Languages* 6 (2006).

[98] Richard Whately. *Elements of Rhetoric: Comprising an Analysis of the Laws of Moral Evidence and of Persuasion: with Rules for Argumentative Composition and Elocution*. Longmans, Green, Reader, and Dyer, 1873.

[99] Yorick Wilks. “A preferential, pattern-seeking, semantics for natural language inference”. In: *Artificial intelligence* 6.1 (1975), pp. 53–74.

[100] Adina Williams, Nikita Nangia, and Samuel Bowman. “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)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 1112–1122. URL: http://aclweb.org/anthology/N18-1101.

[101] Adina Williams, Nikita Nangia, and Samuel R Bowman. “The Multi-Genre NLI Corpus”. In: (2018).

[102] Chuhan Wu et al. “Neural metaphor detecting with CNN-LSTM model”. In: *Proceedings of the Workshop on Figurative Language Processing*. 2018, pp. 110–114.

[103] Ping Xiao et al. “Meta4meaning: Automatic metaphor interpretation using corpus-derived word associations”. In: *Proceedings of the 7th International Conference on Computational Creativity (ICCC). Paris, France*. 2016.

[104] Zhilin Yang et al. “Xlnet: Generalized autoregressive pretraining for language understanding”. In: *Advances in neural information processing systems*. 2019, pp. 5754–5764.

[105] Benat Zapirain et al. “Selectional preferences for semantic role classification”. In: *Computational Linguistics* 39.3 (2013), pp. 631–663.

[106] Omnia Zayed, John P McCrae, and Paul Buitelaar. “Crowd-Sourcing A High-Quality Dataset for Metaphor Identification in Tweets”. In: *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. 2019.

7 Appendix

This task will involve reading an ordered pair of sentences. The senteces will describe a situation or event. Using only this description and what you know about the world classify each sentence pair as an entailment, contradiction, or neutral.

You can use the drop down menu to select which of the three labels you think fits best the current pair.

Assign each of this labels as follows:

- **Entailment:** the second sentence is definitely correct about the situation or event in the first sentence.

Below is an example of an entailment pair:

Sentence 1: "The cottages near the shoreline, styled like plantation homes with large covered porches, are luxurious within; some come with private hot tubs,"

Sentence 2: "The shoreline has plantation style homes near it, which are luxurious and often have covered porches or hot tubs."

- **Neutral:** the second sentence **might** be correct about the situation or event in the first sentence.

Below is an example of a neutral pair:

Sentence 1: "Government Executive magazine annually presents Government Technology Leadership Awards to recognize federal agencies and state governments for their excellent performance with information technology programs,"

Sentence 2: "In addition to their annual Government Technology Leadership Award, Government Executive magazine also presents a cash prize for best dressed agent from a federal agency."

- **Contradiction:** the second sentence is definitely incorrect about the situation or event in the line.

Below is an example of a contradiction pair:

Sentence 1: "Yes, he's still under arrest, which is why USAT's front-page reefer headline British Court Frees Chile's Pinochet is a bit off,"

Sentence 2: "The headline 'British Court Free's Chile's Pinochet' is correct, since the man is freely roaming the streets."

Thank you for your help! If you have more questions, please reach out to the email provided to you.

Figure 7: Instructions Presented to Annotators