A Broader Range for 'Meaning the Same Thing': Human Against Machine on Hard Paraphrase Detection Tasks

Jamie C. Macbeth JMACBETH@SMITH.EDU
Ella Chang ECHANG33@SMITH.EDU
Jingyu Gin Chen JCHEN54@SMITH.EDU
Yining Hua YHUA@SMITH.EDU
Sandra Grandic SGRANDIC@SMITH.EDU
Department of Computer Science, Smith College, Northampton, MA 01063 USA

Winnie X. Zheng WINNIEZ@MIT.EDU
Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139 USA

Abstract

The ability to recognize that pairs or sets of language expressions "mean the same thing" is a cognitive task for which meaning representation is clearly a central issue. This paper uses the task of paraphrasing to study meaning representation in a cognitive system. The main claim of this paper is that a consequential part of the meaning representation for a natural language expression is a set of language-free structures and processes that are not part of the expression in question. To support our claims, we construct a corpus of paraphrase pairs using a system that has a non-linguistic meaning representation decoupled from the linguistic system that generates natural language from it. This corpus of paraphrase pairs is special in that it represents a full range of syntactic and lexical difference in their constituent sentences. We conduct an extensive analysis comparing the performance of a state-of-the-art neural network model against humans performing the paraphrase detection task. We find that the model deviates significantly from human classification performance, particularly on sentence pairs that conveyed the same meaning while exhibiting significant differences lexically and syntactically. As the neural network model is trained only on linguistic items, the discrepancy points to the existence and necessity of a significant non-linguistic part of meaning formation.

1. Introduction

In the mainstream natural language processing AI literature, there are many claims of learned neural network models performing on a par with humans at a variety of natural language understanding and natural language generation tasks. Work on adversarial examples has demonstrated that these systems are not truly understanding language and has created new streams of training examples meant to make the models more robust. However, there is still not a clear and deep understanding of how these celebrated systems actually represent meaning, and whether they are representing meaning in ways similar to humans.

The ability to detect or generate paraphrases, pairs or sets of language expressions that "mean the same thing", while being tasks for which neural systems researchers have claimed success, are clearly cognitive tasks for which meaning representation is a central issue in systems that perform them. The main claim of this paper is that a significant and consequential part of forming a meaning representation of a natural language expression is a set of structures and processes that are not part of the language expressions. A specific part of this claim is that, at least in part, the structures in question are not simply collections of logical assertions that involve the words in the original expression, but cognitive symbolic structures that are language-free—although they correspond to surface language expressions, they do not resemble them and do not necessarily correspond one-to-one with them. We acknowledge that neither of these claims is particularly novel, but we argue they are increasingly relevant in an age where meaning representation is assumed to be happening inside of the state-of-the-art models used in NLP and NLU tasks.

In this paper we perform an extensive analysis of the behavior of a learned neural network trained for the paraphrase detection task as it performs that task on a special corpus of paraphrase pairs. These paraphrase pairs are generated from a system that has a decidedly non-linguistic meaning representation decoupled from a linguistic system that generates natural language from it. They represent a full range of syntactic and lexical difference in their constituent sentences, from paraphrases which are identical except for a single word, to paraphrases that are as different as possible lexically and syntactically. This allowed us to test a hypothesis that the cutting-edge learned neural networks that appear to be able to perform paraphrase detection are not actually understanding sentences in a way similar to humans. We also performed comparisons of our sentence pairs to the sentence pairs in the training and test data for these networks, and performed an experiment with human annotators to test our hypotheses.

We present conclusive evidence that the sentence pairs we have generated are paraphrases. In fact, human participants were more likely to rate them as being fully equivalent in meaning than sentences in a state-of-the-art paraphrasing and semantic similarity task datasets. However, a Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019) model trained on state-of-the-art paraphrasing task datasets deviated significantly from human classification performance on our corpus, particularly on sentence pairs that conveyed the same meaning while exhibiting significant differences lexically and syntactically. Analyzing the training data for the BERT model, we found a significant correlation between surface-level lexical sentence distance measures and human annotations of textual similarity, indicating that the BERT model is trained not to classify sentence pairs based on their similarity in meaning, but largely based on whether they used the same words. We argue that the discrepancy may be related to the methods used to construct the paraphrase corpus, which attempt to align and match sentences from different texts on the same topic. This may indicate a broader challenge to methods and goals of state-of-the-art big-data machine-learning NLP.

2. Background

The term *paraphrase* can be viewed as a relation between surface linguistic forms, a cognitive task of determining that relation, or a cognitive task of generating items for which the relation holds.

In this paper, we focus on the cognitive task of determining, detecting, or recognizing a paraphrase relationship between linguistic items.

Definitions of the paraphrase relation and of paraphrases vary, from defining them as linguistic items, sentences or phrases that "convey the same meaning using different wording" (Bhagat & Hovy, 2013) or "convey almost the same information" (Androutsopoulos & Malakasiotis, 2010) to defining paraphrasing more formally as a textual "meaning-preserving relation" (Culicover, 1968). Madnani & Dorr (2010) cite "the principle of semantic equivalence" and define a paraphrase as "an alternative surface form in the same language expressing the same semantic content as the original form." Some definitions first define "textual entailment,"—a different but somewhat related task of determining that one expression infers another—and then define paraphrase as "bidirectional entailment".

At the very least, we can be sure that paraphrase detection and generation involve meaning of language, and that the complexity and variation of determining and processing meaning, combined with complexities of viewpoint, commonsense and situational inference, and pragmatics, among other factors, are what make posing a precise definition of paraphrasing elusive. This forces a spectrum of "broad" and "strict" views, and "dynamic" or "approximate" definitions such as "quasi-paraphrase", at the same time as it causes some to reject the notion altogether (see Bhagat & Hovy, 2013; Vila et al., 2014). In examining cognitive systems and behavior regarding sentence pair comparison and classification, our theoretical viewpoint is that two major paradigms emerge and are important to define.

2.1 The Surface Alignment Paradigm

Having systems that recognize or generate paraphrases may be used to support a wide range of other major natural language processing applications, such as question answering, query expansion, information extraction, machine translation. But ultimately the utility of paraphrase systems is caused by the fact that computing systems largely lack the powerful cognitive meaning representation and reasoning systems that humans possess and use when processing language. Thus when language is used either as an interface for human users, as a data representation, or in any other aspect of the system, designers must resort to simple mappings between anticipated natural language inputs and system functions. Usability of natural language in these systems is expanded by performing recognition of paraphrase relationships between actual inputs and anticipated inputs, or by generating paraphrases of inputs and attempting to match the texts word-by-word.

One way to approach meaning is to say that it exists entirely within language, and it has no non-linguistic component. Under this theory, meaning consists only of natural language expressions, and paraphrases are simply equivalence relations between lexical and syntactic language forms. One of the main tasks for building systems under this view is the finding and extracting sets of linguistic expressions that comprise the meaning equivalence relation. We name the paradigm for both the extraction of large datasets of paraphrases from large language corpora, as well as the neural models and other machine learning models that are trained on these datasets as operating under the *surface alignment* paradigm.

2.2 The Deep Understanding Paradigm

The fact that we can speak of different language expressions meaning the same thing points to the possibility that there may be a non-linguistic component of meaning representation. However, surface alignment does not access or apply structures of any kind from outside the texts themselves. Another view of cognitive systems and behavior regarding sentence pair comparison and classification postulates that structures and processes that are separate from the language expressions are used to make meaning representations for the expressions, which are then the basis of the comparisons. We name this the *deep understanding* paradigm in contrast to the surface alignment view. We believe that this paradigm is much more akin to how humans and human-like cognitive systems perform sentence pair comparison and meaning formation generally.

In a recent review of literature from diverse perspectives of cognitive science, philosophy, psychology, education, neuroscience, and computer science on understanding writ large, Hough & Gluck (2019) define understanding by focusing on common features such as appropriate use of "organized knowledge structures", "varied or distributed representations" that engender "rich networks of relations," and an incremental process that "often starts at a superficial level and moves towards deeper, more meaningful concepts and relations."

We postulate that these features are also present in understanding of natural language, and that a significant part of forming a meaning representation of a natural language expression is a set of structures and processes that are not part of those expressions; some of these features and representations are not simply logical forms or frames containing words, but are "inner language" forms (Winston, 2012) constructed from symbols of mental imagery (Sadoski & Paivio, 2001). When it comes to paraphrase detection, the systems working under the understanding paradigm process the texts to build rich meaning representations of each language expression using all of the aforementioned processes and attempt to map or match the meaning representations. It could be said that alignment does occur in the deep understanding paradigm as well, but what occurs is an alignment of meaning representations, instead of an alignment of the texts themselves.

3. Methodology

To gather evidence in support of our claims, we built a system representing the deep understanding paradigm. We created a large corpus of paraphrases by using a non-linguistic meaning representation of a concept to generate a set of sentences that use a variety of linguistic expressions to express that meaning. This provided us with a large number of sentence pairs that we hypothesized all meant the same thing and thus were all paraphrases of each other. To represent the alignment paradigm, we used a freely available paraphrase task and dataset created via an alignment procedure. We performed an extensive comparison of the paraphrase corpora, and an analysis of the behavior of a learned neural network trained on the alignment-based corpus as it attempted to discern sentence pairs constructed from the deep understanding paradigm and meaning representation.

Table 1. Example Microsoft Research Paraphrase Corpus (MRPC) paraphrase sentence pair with edit distances and Jaccard distances. Although MRPC sentence pairs are constrained to have a minimum edit distance of 8, Jaccard distances of 0 are possible in sentence pairs that have exactly the same words and punctuation tokens in different order.

Edit Distance	Jaccard Distance	MRPC Paraphrase Sentence Pair
23	0	"'I expect Japan to keep conducting intervention, but the volume is likely to fall sharply,' said Junya Tanase, forex strategist at JP Morgan Chase. " "Junya Tanase, forex strategist at JP Morgan Chase, said 'I expect Japan to keep conducting intervention, but the volume is likely to fall sharply.'"

3.1 Sentence Difference Metrics

Measures of "surface" linguistic similarity between sentences are used both for alignment procedures for creating large paraphrase corpora under the alignment paradigm, and for our comparison studies between alignment-based paraphrase pairs and deep-understanding-based paraphrase pairs. These provide quantitative measures of the syntactic and lexical differences between the sentences in any given sentence pair, and to analyze and compare the sentence pairs from the different datasets we calculated measures of surface similarity between sentences in each pair.

One measure we used was "lexical", word-based edit distance, defined as the minimum number of edit operations (e.g. additions, deletions, and substitutions) of complete words and tokens (e.g. punctuation) needed to transform one sentence into the other. This is a simple measure of both syntactic and lexical difference. Another was Jaccard distance, a measure of lexical dissimilarity between the sets of unique words in each sentence in the pair. It is defined as the difference of the sizes of the union and the intersection of the sets of unique words in the sentences, normalized by the union.

Finally, we used BLEU (the BiLingual Evaluation Understudy) which was originally a scoring metric proposed for evaluating machine translation systems by measuring closeness of a machine translation to that provided by a "professional" translator (Papineni et al., 2002). We use the unigram BLEU-1 score measure, which is a "bags of words" measure of lexical similarity similar to Jaccard distance, but is more sensitive than Jaccard distance because it keeps track of the number of times a word occurs. Given a pair of sentences, called the candidate and the reference, BLEU-1 is a modified unigram precision measure which matches each occurrence of a word in the candidate sentence with an occurrence of the same word in the reference sentence, and divides the count of matches by the total number of words—including duplicates—in the candidate sentence.

3.2 MRPC

For our studies of paraphrasing systems we used the Microsoft Research Paraphrase Corpus (MRPC, Dolan & Brockett, 2005) a well known paraphrase detection and recognition task and part of the General Language Understanding Evaluation (GLUE) benchmark[1]. MRPC consists of 5801 sentence pairs annotated with a paraphrase/non-paraphrase classification. Examples of MRPC sentence pairs are shown in Table 1.

MRPC was constructed through an alignment process that exploited an explosion in Internet news coverage in the early 2000s (Quirk et al., 2004). The first phase of alignment exploited a clustering algorithms used by news aggregator websites to scrape sets of topically- and temporally-related news articles. These articles ideally contain many sentences describing the same events, and have significant overlap in content in reporting the basic facts of a story. A second phase of the alignment attempted to match sentences from articles in a cluster into paraphrase pairs. A lexical edit distance metric was used to compare all pairs of sentences in a news cluster by finding the minimal number of insertions and deletions of words to transform one sentence into the other. Each candidate pair of sentences was required to share at least three words in common, and required to have a lexical edit distance between 8 and 20 edits. Also, the length of the shorter of the two sentences was required to be at least 66.6% that of the longer.

Ultimately 13,127,938 sentence pairs were aligned from 9,516,684 sentences in 32,408 news clusters collected from the World Wide Web over a 2 year period, and then eventually filtered down to 5801 sentence pairs. The sentence pairs were examined by two independent judges who were asked to give a binary judgment on whether the two sentences could be considered "semantically equivalent". A third judge was consulted to resolve disagreements and ultimately 67% of the sentence pairs were judged as paraphrases while 33% were not.

3.3 STS

We also used the Semantic Textual Similarity (STS) dataset, a sample of 1500 sentence pairs from MRPC used as part of the SemEval workshop series on evaluations of semantic analysis. STS has crowdsourced annotations of the MRPC sentence using a more detailed classification system, a ordered, six-point scale indicating "levels" of "semantic similarity",

(5) The two sentences are completely equivalent, as they mean the same thing.

(4) The two sentences are mostly equivalent, but some unimportant details differ.

(3) The two sentences are roughly equivalent, but some important information differs/missing.

(2) The two sentences are not equivalent, but share some details.

(1) The two sentences are not equivalent, but are on the same topic.

(0) The two sentences are on different topics.

1. `https://gluebenchmark.com/`

3.4 BERT

As our exemplar of an alignment-based paraphrase classifier system, we used the Bidirectional Encoder Representations from Transformers deep learning system (BERT, Devlin et al., 2019). BERT is known as a "general purpose architecture for natural language understanding" and it achieves improved performance scores over state-of-the-art models on a number of natural language processing tasks and datasets.

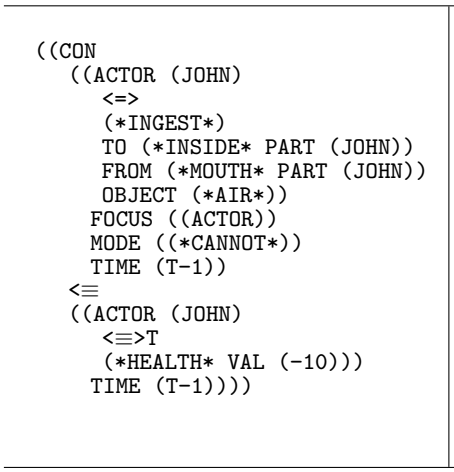
BERT's transformer sequence-to-sequence model architecture allows for efficient pre-training of language models over vast unlabeled language corpora through parallelization, followed by a second "fine-tuning" training phase on a smaller, labeled dataset for a specific task. BERT's distinctive advances in neural architectures for NLP were its elimination of unidirectionality constraints, its greater ability for learning long-range relationships in text easier to learn, and its presentation of a unified architecture capable of a variety of language processing tasks, for classification as well as language generation. For our studies we used an instance of a version of BERT called $\text{BERT}_{\text{BASE}}$ that had been fine-tuned to perform the MRPC classification task. $\text{BERT}_{\text{BASE}}$ has 110M total parameters.

3.5 BABEL

As our representative of the deep understanding paradigm, we built a system to generate a considerably large number of paraphrases of the same idea. The system is based on a natural language generation and paraphrasing system called BABEL (Goldman, 1975), part of the Memory, Analysis, Response Generation, and Inference in English system (MARGIE, Schank et al., 1975) a classic demonstration of meaning representation featuring non-linguistic structures.

BABEL generates natural language surface realizations of structures represented by language-free conceptual base called Conceptual Dependency (CD, Schank, 1972) using an augmented transition network, or ATN (Simmons & Slocum, 1972). BABEL first runs the CD structure through a discrimination net, which selects a matching word sense. Each word sense has a corresponding entry in the BABEL conceptual lexicon (called the "concexicon"), which carries both an ATN grammar symbol corresponding to the syntactic context in which the word sense can be generated and information about how the word sense is generated grammatically. In cases when the word sense may have subordinate clauses or phrases, the concexicon contains their corresponding ATN grammar symbols, as well as pointers to the substructures of the CD structure containing the conceptual information for generating these subordinates. A semantic network node is created corresponding to the word sense, and the procedure is called recursively on the subordinate CD structures. The semantic network node of the head word sense is furnished with links to semantic network nodes generated by recursive calls. Once the semantic network generation process is completed, the ATN generation procedure applies an English grammar to the semantic network starting with the top-level node to generate a surface-level realization in the form of an English sentence.

We ran BABEL in its AND-OR paraphrasing mode to generate every possible paraphrase of a CD structure. With significant modifications and additions to the original BABEL conceptual lexicon and ATN grammar, we generated 432 sentences, from which all possible matchings gave us



```
((CON
   ((ACTOR (JOHN)
      <=>
      (*INGEST*)
      TO (*INSIDE* PART (JOHN))
      FROM (*MOUTH* PART (JOHN))
      OBJECT (*AIR*))
     FOCUS ((ACTOR))
     MODE ((*CANNOT*))
     TIME (T-1))
   <≡
   ((ACTOR (JOHN)
      <≡>T
      (*HEALTH* VAL (-10)))
     TIME (T-1))))
```

"John died because he could not breathe."
"The cause of the end of John 's life was his inability to take a breath."

"Not being able to inhale air made John die."
"John 's life ended because he could not take a breath."

"John became dead because he could not inhale air."
"John 's death resulted from him being unable to breathe."

Figure 1. Left: the Conceptual Dependency structure s-expression used to generate sentences with BABEL. The INGEST primitive act is a decomposition of breathing, while *CANNOT*, <≡, and <≡>T represent disablement, causation, and state change primitives respectively. Right: example BABEL paraphrase sentence pairs. Each sentence pair has Jaccard distance greater than 0.94.

93,096 sentence pairs for a paraphrase corpus. The CD structure used and examples of the BABEL sentences are shown in Figure 1.

3.6 Human Subject Studies

We also performed a study with human participants as representatives of the understanding paradigm of paraphrase recognition in action. 208 participants took part in the study through an Amazon Mechanical Turk human intelligence task (HIT). Participants were "Masters" Turk workers with a 90% approval rate and at least 1000 previously approved HITs.

The survey presented each participant with a BABEL sentence pair and an MRPC sentence pair. For each pair of sentences, the participant was required to rate the sentence pair on the six-point STS scale to indicate how similar the sentences were in meaning. We also included a third sentence pair as an "attention check" item to ensure that participants were actively reading the items and response options and not simply answering randomly. The attention check items were created by sampling sentences from entirely different datasets in the GLUE benchmark and manually matching them so that the sentences in the pair were always obviously on different topics. Submissions from participants who did not classify these items as 0 on the STS scale were rejected and discarded. The ordering of the sentence pairs was counterbalanced to eliminate ordering effects in participants' answers.

The 208 BABEL sentence pairs which were presented to participants were sampled from the full BABEL corpus in three different groups. In one group, 100 BABEL sentence pairs were selected evenly spaced over the full range of Jaccard distances from 0 to 1. For a second group we chose the 68 sentence pairs which had the maximum Jaccard distance of 0.957. In the third, we chose the 40 sentence pairs which had BLEU score ≤ 0.4. We chose these groups to assure that we had

significant sampling of the humans' behavior for sentence pairs with large distances. For the MRPC sentence pairs, we randomly sampled 208 sentence pairs from the MRPC training set which were marked as paraphrases in the original MRPC annotation.

4. Results

4.1 Sentence Length and Distance Measures

4.1.1 Sentence Lengths

Before comparing sentence pairs based on distance measures, we wanted to make sure that they were comparable in terms of their lengths. We calculated the number of word-tokens in all of the MRPC sentences. They ranged from 5 to 31 tokens and had a mean of 18.9 tokens. The BABEL sentences were typically shorter; they had a mean of 12.3125 tokens per sentence, and ranged from 7 tokens to 18 tokens.

4.1.2 Jaccard and Edit Distances

We calculated lexical edit distances as a mixed measure of both lexical and syntactic similarity for all of the sentence pairs in our study. The MRPC sentences had minimum and maximum edit distances of 4 and 28 respectively, while the mean edit distance was 11.57. Since MRPC has both paraphrase and non-paraphrase pairs, we isolated the paraphrase sentence pairs. They had a mean edit distance of 10.79. Because the BABEL sentences tended to be shorter, it is understandable that their range of edit distances, from 1 to 18, was also smaller. In spite of this, the median and mean of edit distance for the BABEL sentence pairs was on par with that of the MRPC sentences.

We also calculated lexical Jaccard distance for all sentence pairs, which represents lexical dissimilarity as the ratio of intersection over union for the bags of words in each sentence in the pair. For the 5801 MRPC sentence pairs the minimum and maximum Jaccard distances were 0 and 0.846 and the mean Jaccard distance was 0.481. For the MRPC sentences which were annotated as paraphrases, the mean Jaccard distance was 0.437. There were several sentence pairs with Jaccard distance of 0, indicating that they used exactly the same words. An example of an MRPC sentence pair with 0 Jaccard distance is shown in Table 1

The BABEL sentences had a Jaccard distance mean of 0.7 and a median of 0.72, ranging from 0 to 0.957. In spite of the fact that a large percentage of the MRPC sentences are classified as non-paraphrases, the BABEL sentence pairs exhibited a greater degree of lexical difference. We performed a Mann-Whitney U test comparing the Jaccard distances as an interval dependent variable for the BABEL sentence pairs and the MRPC sentence pairs and found that the Jaccard distance to be significantly larger statistically for the BABEL sentences ($U \simeq 2.0 \times 10^8$, $p < .001$).

4.2 Human Classifications and BERT Classifications

4.2.1 Sentence Pair Annotations

Human evaluators found the 208 BABEL sentence pairs to be "completely equivalent" (5 on the STS scale) 92.8% of the time and "mostly equivalent" (4 on the STS scale) an additional 6.2% of the

Table 2. Semantic textual similarity scores on BABEL and MRPC sentence pairs from a human participants study. BABEL-Max refers to the subset of BABEL sentence pairs with the greatest Jaccard distance.

	(5) "completely equivalent"	(4) "mostly equivalent"	(3) "roughly equivalent"	(2) "share ... details"	(1) "same topic"	(0) "different topics"
MRPC	30.8%	45.6%	21.2%	2.4%	-	-
BABEL	92.8%	6.2%	0.5%	-	0.5%	-
BABEL-Max	94.1%	5.9 %	-	-	-	-

time with a mean STS scale score of 4.9. Participants only rated the BABEL sentences < 4 in two cases.

In contrast, human evaluators found the 208 MRPC sentences that were classified as paraphrases by the original MRPC annotators to be "completely equivalent" only 30.8% of the time. They actually labeled a higher percentage (45.6%) as "mostly equivalent" and more than 1/5th as "roughly equivalent", with a mean STS score of 4.05, and a median of 4. As a result, we can conclude confidently that the BABEL sentences pairs truly are paraphrases, or at least they can be considered paraphrases at least as much as the MRPC sentences can be.

When we ran a BERT model trained on the MRPC dataset on the BABEL sentence pairs, it classified 82.1% of them as paraphrases. This is in comparison to an accuracy of 88.2% that BERT is reported to achieve on MRPC. What follows, however, is a deeper analysis showing that BERT is it highly dependent on surface similarity of the sentence pairs in making its judgments.

4.3 Correlation Measures

4.3.1 Correlations between Distance Measures and Bert Classifications

To determine the degree to which BERT may be detecting paraphrases largely by comparing whether the two sentences in the pair use the same words, we calculated the percentage of BABEL sentences that BERT classified as paraphrases for subsets of the BABEL sentence pairs based on their distance measures.

Unsurprisingly, at the bottom end of the range for Jaccard distance, where the sentences are nearly the same, BERT classifies 100% of the pairs as paraphrases. However, the BERT percentage on the BABEL sentences declined to 80.1% when only considering sentence pairs with edit distance > 8. In particular, for the 68 BABEL sentence pairs which had the greatest Jaccard distance of 0.957, BERT classified them as paraphrases only 61.7% of the time. In contrast, human raters' mean STS rating of the 68 BABEL sentence pairs which had the greatest Jaccard distance of 0.957 was 4.94. They found them to be "completely equivalent" 94.1% of the time and "mostly equivalent" an additional 5.9% of the time.

To determine whether the human evaluators were also biased toward classifying the BABEL sentence pairs as paraphrases by surface-level distance measures, we plotted the STS ratings given

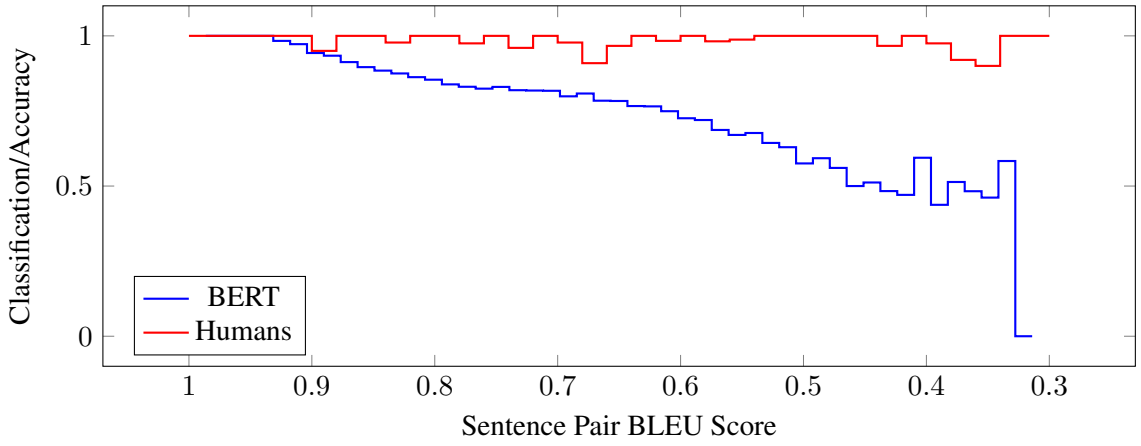


Figure 2. Accuracy of BERT and human annotators for BABEL sentence pairs as a function of BLEU score for the pair. Smaller values of BLEU indicate less lexical overlap or greater lexical "distance". Human classifications are STS classifications (0-5) that have been scaled to the 0-1 classification range of BERT and the MRPC data, with STS values of 5 ("completely equivalent") scaled to 1. The plot represents the average of classifications over bins that were equally-sized over the range of BLEU scores. The trend in BERT classifications indicates that lexical similarity is a major cue in BERT's classification behavior, while the human annotators use a meaning representation of the sentences to make their determinations.

by human evaluators on the BABEL sentence pairs as a function of their BLEU distances against their BERT classifications. Against these we also plotted BERT's classifications of the BABEL sentence pairs, also as a function of their BLEU distances (Figure 2). As with Jaccard distance, the accuracy is high for small distances, but declines consistently over the range of BLEU values, falling to roughly 50% accuracy about three-quarters into the range. This is a dramatic divergence from the behavior of human raters, who classify the BABEL sentences as paraphrases at the same rate throughout the range (Figure 2).

We also calculated whether BERT's classifications were statistically correlated to the Jaccard distance or edit distance between the sentences in the pairs. We performed a Mann-Whitney U test comparing the means of the Jaccard distances as interval dependent variables for the pairs of BABEL sentences that BERT classified as paraphrases and those that BERT classified as non-paraphrases as two independent groups. We found a statistically significant difference ($U \simeq 1.95 \times 10^9$, $p < .001$) between these, an indication that BERT is examining the words in the sentences instead of their meanings to perform the classification. We performed the same test for lexical edit distance and again found a statistically significant difference ($U \simeq 2.24 \times 10^9$, $p < .001$).

4.3.2 Correlations between Distance Measures and Annotations in MRPC

To determine the degree to which BERT's behavior on the BABEL paraphrases is a reflection of the training data, we calculated a Mann-Whitney U test comparing the Jaccard distances for the 67% of the original 5801 MRPC pairs that were annotated as paraphrases against the 33% which

were not. We found that both Jaccard distances and edit distances were statistically larger for the non-paraphrase pairs in MRPC, but difference in means was not dramatic: the mean edit distance of the MRPC paraphrase pairs was 10.8, while for non-paraphrase pairs it was 13.2; the mean Jaccard distance of paraphrase pairs was 0.437 but was 0.57 for non-paraphrase pairs.

The 1500 sentences in the Semantic Textual Similarity dataset (Agirre et al., 2012) are MRPC sentences which are annotated with "gold standard" STS scores—for each sentence pair, the annotation is an average of 5 scores given by crowdworkers. We calculated a Spearman's ρ correlation between Jaccard distance and the gold standard score in the STS dataset expecting lower gold standard similarity scores for pairs with greater Jaccard distance. We found a statistically significant correlation between Jaccard distance and the gold standard score ($r_s = -0.494$, $p < 0.001$) and a significant correlation between edit distance and gold standard score ($r_s = -0.115$, $p < 0.001$). We can conclude that BERT's bias toward classifying lexically similar sentences as paraphrases may be due to its training.

4.3.3 Lack of Correlations between Distance and Human Classifications

We saw in Figure 2 that human raters appear to classify the BABEL sentences as paraphrases at the same rate throughout the range of BLEU lexical distances, and were not affected by the high degree of lexical dissimilarity in detecting that sentence pairs were paraphrases. In order to provide statistical verification of this assumption, we ran a Spearman's ρ test for a correlation between human ratings of the BABEL sentences and Jaccard distance. Based on the test results, we were unable to invalidate the null hypothesis that they are not correlated ($r_s = 0.03$, $p = 0.6$). Similar tests for correlations between human ratings and edit distance and BLEU score produced similar results and also did not invalidate the null hypotheses. We can conclude that there is no such correlation and that the skill that human raters have at detecting that a two sentences are paraphrases is independent of whether the sentences in the pair are lexically or syntactically similar.

5. Discussion

Our results provide a clear indication of the differences in paraphrase recognition behavior, and, consequently, meaning representation behavior between humans and the BERT/MRPC model. At the core of its design, BERT is a language model, and it is a prime example of a system that manipulates language, but does not have a meaning representation system apart from sequences of language symbols found in corpora. The deep understanding theory of non-linguistic meaning representation appears to result in systems that more accurately resemble human language behavior.

We also argue for our claims based on an examination of proposals for how the paraphrasing task is correctly performed under the different paradigms. Under the surface alignment paradigm, the paraphrase detection presumably works by direct matching of linguistic expressions, which alone form the substrate for the paraphrase relation. For example, one possibility is that humans are matching at a sentence-wide level, and storing separate relation entries for the 93K BABEL sentence pairs and any other possible sentence pair that could be encountered in the task. However, this option seems impossible given the uncountably infinite nature of language production; one would need a pre-determined stored relation for any possible unheard sentence.

We can also argue against the proposal that humans are storing direct equivalence relations between words, phrases, or smaller linguistic units due to the inefficiency of doing so, since an equivalence between n expressions would require $O(n^2)$ relation entries. Establishing equivalence through chains of relations in which a particular linguistic expression serves as a representative of a set does cut down on this expense, but this solution begins to resemble the separate, non-linguistic meaning representation we propose in our claim. We can likewise argue against proposals that human understanders process language into logical formulas, since, other than logical operators, the logical formulas will have words and linguistic expressions as propositions and predicates, and thus equivalence may engender the same proliferation of relations as with language.

On the other hand, under the deep understanding paradigm, if language inputs are transformed to a representation based on an interlingual conceptual base, this base form acts as a representative of the equivalence relation, which can then be implemented in linear space. We can support this argument subjectively by citing the way that, in our experience, BABEL's decoupling of linguistic knowledge from non-linguistic CD conceptual representations facilitates the generation of so many paraphrases combinatorically with relatively little input.

But it is also important to consider the natural language understanding tasks that researchers are focused on when training and testing models like BERT and whether the training datasets express a full range of human language behaviors. We found that not even a majority of the MRPC sentences which were originally annotated as paraphrases were considered "completely equivalent" on the STS scale by our participants. This seems to be a failure of the alignment method for creating paraphrase datasets.

In the big-data NLP literature on textual similarity and paraphrase detection, the main clues pointing to the challenges of the alignment method as a representation of human language behavior are in the evolution of the classification systems themselves, which retreat from expressions of meaning equivalence to embrace vague gradations of "semantic relatedness" and "meaning overlap". Dolan & Brockett (2005) admit this issue, stating that sentences judged 'semantically equivalent' in MRPC "in fact diverge semantically to at least some degree" and have "obvious differences in information content".

One explanation for this is that the original method for constructing MRPC attempts to align sentences from a cluster of news stories on a particular event. Firstly, many news stories are only "minor edits" of an original AP or Reuters story. Also, different news stories may express the same details, but the writers may form sentences using different combinations of those details, such that no pairing of sentences from one story with sentences from another results in a paraphrase pair with equivalent meaning. As a result, alignments may almost never find pairs of sentences that express nearly exactly the same idea and that vary significantly lexically and syntactically with out varying significantly in the details that are expressed.

6. Related Work

Androutsopoulos & Malakasiotis (2010) survey paraphrasing and textual entailment methods, and their discussion does cover recognition, generation, and extraction of paraphrases, as well as non-data-driven symbolic methods. Madnani & Dorr (2010) also survey data-driven alignment-based

paraphrase generation. Bhagat & Hovy (2013) attempt to avoid conflict between strict and broad notions of paraphrasing to define paraphrases as "quasi-paraphrases" conveying "approximately the same meaning" while performing a manual examination of MRPC.

Over the past decade, a variety of neural architectures have been developed for paraphrasing tasks. Recursive autoencoders (Socher et al., 2011), convolutional neural networks (Yin & Schütze, 2015; He et al., 2015), recursive neural networks (Chen et al., 2018), and long short-term memory architectures (Lan & Xu, 2018) have all been trained to perform paraphrase identification and sentence similarity modeling. Many of these architectures have also been used for paraphrase generation (Prakash et al., 2016; Li et al., 2018). Other research has probed the internal representations of BERT and other language-oriented neural networks (Tenney et al., 2019; Blevins et al., 2018), but this work is geared toward quantifying how syntax and surface-level linguistic forms are captured by these models, not how they represent meaning.

There are various examples in the literature of alignment-style paraphrase sentence pair extraction. Lin & Pantel (2001) confront alignment from an information extraction perspective, while Barzilay & McKeown (2001) align sentences in multiple English translations of classic literary works. Barzilay & Lee (2003) calculate an alignment form called "lattices"—graph-based representations of structural similarity in sentences in a corpus of news articles—and pair them together as paraphrases if they tend to have the same entities as arguments. Pang et al. (2003) extract synonyms and phrasal paraphrases from multiple translations of news articles as finite state automata (FSAs) and perform a human evaluation, but only of synonyms and short phrasal paraphrases represented by their FSAs.

Much of the legacy work on non-alignment-based paraphrase generation uses hand-crafted rules to construct a syntax-focused representation of the input, and to generate paraphrases based on it. Sparck Jones & Tait (1984) use a linguistically-motivated generator to provide a systematic range of variant expressions of indexing concepts for document retrieval, while McKeown (1979) generates paraphrases of natural language database queries to aid infrequent users of database systems, using syntactic hand-crafted methods. These systems, however, are limited to shallow syntactic representations of meaning.

7. Conclusion

In this paper we worked to support claims about the character of meaning representation systems required to perform cognitive tasks related to natural language understanding. We argued that symbolic structures and processes that are disparate from the language expressions are required to make meaning representations for those expressions. We studied the cognitive tasks of paraphrase recognition and generation to provide evidence supporting these claims.

We performed an extensive analysis of the behavior of a learned neural network trained on a large corpus of paraphrases based on an alignment model for pairing sentences, which corresponds to purely linguistic theories of meaning. We ran the neural paraphrase detector on a set of paraphrase pairs that we generated from a non-linguistic meaning representation exhibiting a full range of syntactic and lexical difference in their constituent sentences, and found that it deviated significantly from human classification performance, particularly on the sentence pairs that had significant differ-

ences linguistically. These differences imply that either the neural models or the datasets that they are trained on are missing this fundamental component of meaning formation and representation.

Work on adversarial examples for neural networks, rather than convincing researchers of the drawbacks of research paradigms that focus on shallow performance metrics of models on big datasets, has recently been subverted into another method for generating more training data that ostensibly improves models and makes them more robust. We resist the urge to turn sentence pairs created by BABEL into data for BERT's fine-tuning for paraphrase detection tasks. Although our study shows that BERT's behavior may only be a reflection of weaknesses in the MRPC, we think a more important avenue of future investigation would be to somehow train a model that represents the non-linguistic aspects of meaning present in BABEL. This will be challenging because there is no known dataset of the size that would make this feasible for a neural model, and even if this is achieved, using this model to process or generate text will at least require training data for BABEL's semantic network system, which builds linguistic surface realizations via a mapping from the non-linguistic meaning representations.

The accomplishments of this paper were only possible because we expended the effort to build a symbolic system by hand as an expression of a meaning representation theory, instead of focusing on building a large dataset. While some AI traditions have grown to fear building these kinds of systems because of the "knowledge bottleneck" and the impractical prospect of scaling a symbolic system by hand into a complete human-level intelligence, our future work is not discouraged by these prospects, because it focuses on using hand built systems as simulation tools to test cognitive system theories at a small scale which can later be used to enhance or complement big-data and deep learning approaches. Since models like BERT do not appear to lead to a better understanding of what human language understanding is, we intend to investigate other natural language understanding datasets and models through the kind of study presented in this paper, while exploring improvements in the non-linguistic meaning representation system.

References

Agirre, E., Diab, M., Cer, D., & Gonzalez-Agirre, A. (2012). SemEval-2012 task 6: A pilot on semantic textual similarity. *Proceedings of the Sixth International Workshop on Semantic Evaluation* (pp. 385–393). Montréal, Canada: Association for Computational Linguistics.

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, *38*, 135–187.

Barzilay, R., & Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 16–23). Edmonton, Canada: Association for Computational Linguistics.

Barzilay, R., & McKeown, K. R. (2001). Extracting paraphrases from a parallel corpus. *Proceedings of the Thirty-Ninth Annual Meeting of the Association for Computational Linguistics* (pp. 50–57). Toulouse, France: Association for Computational Linguistics.

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

Blevins, T., Levy, O., & Zettlemoyer, L. (2018). Deep RNNs encode soft hierarchical syntax. *Proceedings of the Fifty-Sixth Annual Meeting of the Association for Computational Linguistics* (pp. 14–19). Melbourne, Australia: Association for Computational Linguistics.

Chen, Q., Hu, Q., Huang, J. X., & He, L. (2018). CA-RNN: using context-aligned recurrent neural networks for modeling sentence similarity. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 265–273). New Orleans, Louisiana, USA: AAAI Press.

Culicover, P. W. (1968). Paraphrase generation and information retrieval from stored text. *Mechanical Translation and Computational Linguistics*, *11*, 78–88.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Dolan, B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. *Proceedings of the Third International Workshop on Paraphrasing*. Jeju Island, South Korea: Asia Federation of Natural Language Processing.

Goldman, N. M. (1975). Sentence paraphrasing from a conceptual base. *Communications of the ACM*, *18*, 96–106.

He, H., Gimpel, K., & Lin, J. (2015). Multi-perspective sentence similarity modeling with convolutional neural networks. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 1576–1586).

Hough, A. R., & Gluck, K. A. (2019). The understanding problem in cognitive science. *Advances in Cognitive Systems*, *8*, 13–32.

Lan, W., & Xu, W. (2018). Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3890–3902).

Li, Z., Jiang, X., Shang, L., & Li, H. (2018). Paraphrase generation with deep reinforcement learning. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3865–3878). Brussels, Belgium: Association for Computational Linguistics.

Lin, D., & Pantel, P. (2001). Discovery of inference rules for question-answering. *Natural Language Engineering*, *7*, 343–360.

Madnani, N., & Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, *36*, 341–387.

McKeown, K. R. (1979). Paraphrasing using given and new information in a question-answer system. *Proceedings of the Seventeeth Annual Meeting of the Association for Computational Linguistics* (pp. 67–72). La Jolla, California, USA: Association for Computational Linguistics.

Pang, B., Knight, K., & Marcu, D. (2003). Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational*

Linguistics (pp. 181–188). Edmonton, Canada: Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the Fortieth Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). Philadelphia, PA: Association for Computational Linguistics.

Prakash, A., Hasan, S. A., Lee, K., Datla, V., Qadir, A., Liu, J., & Farri, O. (2016). Neural paraphrase generation with stacked residual LSTM networks. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2923–2934). Osaka, Japan: The COLING 2016 Organizing Committee.

Quirk, C., Brockett, C., & Dolan, W. (2004). Monolingual machine translation for paraphrase generation. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* (pp. 142–149). Barcelona, Spain: Association for Computational Linguistics.

Sadoski, M., & Paivio, A. (2001). *Imagery and text: A dual coding theory of reading and writing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schank, R. C. (1972). Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology*, *3*, 552–631.

Schank, R. C., Goldman, N. M., Rieger III, C. J., & Riesbeck, C. K. (1975). Inference and paraphrase by computer. *Journal of the ACM*, *22*, 309–328.

Simmons, R., & Slocum, J. (1972). Generating English discourse from semantic networks. *Communications of the ACM*, *15*, 891–905.

Socher, R., Huang, E. H., Pennin, J., Manning, C. D., & Ng, A. Y. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in Neural Information Processing Systems* (pp. 801–809).

Sparck Jones, K., & Tait, J. I. (1984). Automatic search term variant generation. *Journal of Documentation*, *40*, 50–66.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *Proceedings of the Fifty-Seventh Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Florence, Italy: Association for Computational Linguistics.

Vila, M., Martí, M. A., & Rodríguez, H. (2014). Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, *4*, 205–218.

Winston, P. H. (2012). The right way. *Advances in Cognitive Systems*, *1*, 23–36.

Yin, W., & Schütze, H. (2015). Convolutional neural network for paraphrase identification. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 901–911).