Advances in Cognitive Systems 10 (2023) 85–102 Submitted 5/2020; published 7/2023

Humans Against Large Language Models 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 is that a consequential part of the meaning representation for a natural language expression is a set of language-free structures that are not part of the expression in question. To support this claim, 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 its constituent sentences. We conduct an extensive analysis that compares the performance of a neural network model and humans on the paraphrase detection task. We find that, unlike humans, the model fails to recognize paraphrases when the sentences use different words and syntactic structures to convey the same meaning. As the neural network model is trained only on linguistic items, the discrepancy points to the existence of a substantial 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 generation tasks. However, research on adversarial examples has demonstrated that these systems do not truly understand language and that neural systems researchers may have prematurely claimed success on these types of tasks (Iyyer et al., 2018). It remains unclear how well these celebrated systems actually represent meaning and whether they encode meaning in ways similar to humans.

The main claim of this paper is that a substantial and consequential part of forming a meaning representation of a natural language expression is the processing of structures that are not themselves language expressions. A specific part of this claim is that the structures in question are not simply

collections of logical assertions that involve the words in the original expression, but rather 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 the neural models used on natural language processing and understanding tasks.

The tasks of detecting and generating paraphrases—pairs or sets of language expressions that "mean the same thing"—are clearly cognitive tasks for which meaning representation is a central issue. In this paper we perform an extensive analysis of the behavior of a learned neural network on a special corpus of paraphrase pairs generated by a system that uses non-linguistic meaning representations. This corpus represents a full range of syntactic and lexical difference in its constituent sentences, from paraphrases that are identical except for a single word to paraphrases that are as different as possible lexically and syntactically. This let us test a hypothesis that the current learned neural networks for paraphrase detection do not actually understand sentences in a way similar to humans. We also compare our sentence pairs to those used to train and test these networks, employing human annotators to test our hypotheses.

We present conclusive evidence that our sentence pairs are paraphrases. In fact, human participants were more likely to rate them as being fully equivalent in meaning than pairs in data sets widely used for semantic similarity tasks. However, a Bidirectional Encoder Representations from Transformers (BERT, Devlin et al., 2019) model trained on paraphrase recognition data sets deviated substantially 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 correlation between surface-level lexical sentence distances and human annotations of textual similarity. This indicates that the BERT model is trained to classify sentence pairs not based on their similarity in meaning but whether they use the same words. We argue that this discrepancy may be related to the methods used to construct the paraphrase corpus, which align and match sentences from different texts on the same topic. This may indicate a broader challenge to the methods and goals of statistical language processing.

2. Background

The term *paraphrase* can refer to 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 task of determining, detecting, or recognizing a paraphrase relationship among linguistic items. Definitions of the paraphrase relation and of paraphrases vary, from defining them as 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 and 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 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 the meaning of language. However, the complexities of viewpoint, commonsense and situational inference, pragmatics, and other factors that cause variation in how meaning is determined make posing a precise definition of paraphrasing elusive. This forces a spectrum of "broad" and "strict" views alongside "dynamic" or "approximate" definitions such as "quasi-paraphrase", while causing some to reject the notion altogether (e.g., Bhagat & Hovy, 2013; Vila et al., 2014). In the AI literature on sentence pair comparison and classification, two major paradigms emerge that merit closer inspection.

2.1 The Surface Alignment Paradigm

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, and machine translation. But ultimately the utility of paraphrase systems is because computing systems largely lack the powerful representation and reasoning abilities that humans use when processing language. Thus, when language is used as an interface for human users, designers must specify simple mappings between anticipated language inputs and system functions. One expands these systems' coverage by extending paraphrase relationships between actual inputs and anticipated inputs, or by generating paraphrases of inputs and attempting to match the paraphrases with the anticipated inputs 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. A primary task for building systems under this view is finding and extracting sets of linguistic expressions that comprise the meaning equivalence relation. We will use the term *surface alignment* to refer to the extraction of data sets of paraphrases from large language corpora and the operation of learned models that are trained on these data sets. The vast majority of data sets for such training for paraphrase detection focus on some form of surface alignment (e.g., Dolan & Brockett, 2005; Lin & Pantel, 2001; Barzilay & McKeown, 2001; Barzilay & Lee, 2003; Pang et al., 2003)

2.2 The Deep Understanding Paradigm

The fact that we can speak of different sentences or phrases meaning the same thing suggests that there may be a non-linguistic component of meaning. However, systems working under the surface alignment paradigm do not access or apply structures of any kind from outside the texts themselves. Another approach to comparing sentence pairs postulates that the structures and processes used to make meaning representations are separate from language. These meaning representations then provide the basis for comparing and classifying the sentence pairs. We claim that this *deep understanding* paradigm is much more akin to how humans actually process meanings.

In a recent review of literature from diverse perspectives of cognitive science, philosophy, psychology, education, neuroscience, and computer science on understanding writ large, Hough and 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 major part of extracting meaning from a natural language expression involves structures and processes that are not part of the expression. 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, systems in the understanding paradigm process the texts to build rich meaning representations of each language expression and attempt to map or match the meaning representations. We might say that alignment occurs in the deep understanding paradigm as well, but that meanings are aligned rather than the texts themselves.

3. Hypotheses, Data Sets, and Metrics

To gather evidence in support of our claim that non-linguistic representations are key to human understanding, we conducted studies of surface alignment and deep understanding systems. We hypothesized that systems based on the alignment paradigm largely detect paraphrases by examining the surface language features—the words and syntactic constructions—of the sentence pair, while deep understanding systems employ non-linguistic representations to detect paraphrases. We operationalized this informal hypothesis with an experiment that tests whether an alignment-based system can detect paraphrases as easily as a deep understanding system when the two sentences in a paraphrase pair are not aligned lexically and syntactically. This section describes the data sets, models, metrics, and overall methodology of our study.

3.1 Sentence Difference Metrics

To facilitate an experiment that tests whether an alignment-based system detects paraphrases when the two sentences in a paraphrase pair are not aligned lexically and syntactically, we computed quantitative measures of the syntactic and lexical differences between the sentences in sentence pairs and used them as an independent variable in the paraphrase detection experiment. We used the same measures of "surface" linguistic similarity that were used for alignment procedures for creating the paraphrase corpora under the alignment paradigm, and we added an extra measure for its greater sensitivity.

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 the Jaccard distance, a measure of lexical dissimilarity between the sets of unique words in each sentence in the pair. This is defined as the difference in size of the union and intersection of unique word sets in the sentences, normalized by the former.

Finally, we used BLEU (the BiLingual Evaluation Understudy), which was originally a scoring metric proposed for evaluating machine translation systems in terms of closeness to the translation provided by a "professional" translator (Papineni et al., 2002). Specifically, we used the unigram BLEU-1 score measure, which is a "bag of words" measure of lexical similarity similar to Jaccard distance, but is more sensitive because it tracks how often a word occurs. Given a pair of sen-

Table 1. Example Microsoft Research Paraphrase Corpus (MRPC) paraphrase sentence pair with the edit distance and Jaccard distance between the sentences. Although MRPC sentence pairs are constrained to have a minimum edit distance of eight, Jaccard distances of zero are possible in sentence pairs that have exactly the same words and punctuation tokens in a 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.'"

tences, called the candidate and the reference, BLEU-1 is a modified unigram precision measure that matches each occurrence of a word in the candidate with an occurrence of the same word in the reference and divides the count of matches by the total number of words—including duplicates—in the candidate sentence.

3.2 Paraphrase Data Sets

For our experiment, we used two existing paraphrase corpora that correspond to the alignment paradigm and we created a novel paraphrase data set to represent the deep understanding paradigm. This section describes each of these data sets.

3.2.1 The Microsoft Research Paraphrase Corpus

To represent the alignment paradigm, 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 5,801 sentence pairs annotated with a paraphrase/non-paraphrase classification. Table 1 shows one example of an MRPC sentence pair.

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 clustering algorithms used by news aggregator Web sites to scrape sets of topically- and temporally-related news articles that contain 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 and to have a lexical edit distance between 8 and 20. Also, the length of the shorter sentence was required to be at least 66.6% that of the longer one.

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 two year period, and then eventually filtered

1. https://gluebenchmark.com/

to 5,801 sentence pairs. These pairs were examined by two independent judges who gave a binary judgment on whether the two sentences could be considered "semantically equivalent". A third judge was consulted to resolve disagreements and the judges collectively annotated 67% of the sentence pairs as paraphrases and 33% as non-paraphrases.

3.2.2 Semantic Textual Similarity

To enhance our representation of the alignment paradigm, we also used the Semantic Textual Similarity (STS) data set, a sample of 1,500 sentence pairs from MRPC used as part of the SemEval workshop series on evaluations of semantic analysis (Agirre et al., 2012). STS defines a classification system based on an ordered, six-point scale that indicates "levels" of "semantic similarity". As given in Agirre et al. (2012), the points of the scale are:

(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.

The STS data set contains additional crowdsourced annotations of the MRPC sentence pairs using this more detailed classification system. We also used this scale in our main experiment.

3.2.3 BABEL

To represent the deep understanding paradigm, we created a third, novel data set by building a system to generate several hundred sentences, with each being a surface realization of the same non-linguistic conceptual structure. This made every sentence a paraphrase of every other sentence in this set, allowing us to create a corpus of paraphrases that consisted of every possible pairing of sentences in the set.

The system that generated the sentences is based on a component of the Memory, Analysis, Response Generation, and Inference in English system (MARGIE, Schank et al., 1975), a classic demonstration of meaning representation with non-linguistic structures. MARGIE had three components: a conceptual analyzer that maps natural language into conceptual structures, an engine for making inferences from the conceptual structures (called MEMORY), and a generator that maps conceptual structures back into natural language (˙BABEL, Goldman, 1975). We used only BABEL, the natural language generation component. On its own, BABEL cannot understand language or recognize paraphrases, but it can generate a combinatorially large number of sentences from a single conceptual structure. We used this feature to construct a corpus of paraphrase pairs with a broad range of syntactic and lexical difference in their constituent sentences.

BABEL uses an augmented transition network (Simmons & Slocum 1972; Woods, 1970) to generate natural language surface realizations of structures represented by a language-free conceptual base called Conceptual Dependency (CD, Schank, 1972). It first runs the CD structure through a discrimination net, which selects a matching word sense. Each sense has a corresponding entry in the conceptual lexicon, called the "concexicon", which carries both a grammar symbol correspond-

Table 2. Left: The Conceptual Dependency structure s-expression used to generate sentences with BABEL. The INGEST primitive is a representation of breathing, while *CANNOT*, <≡, and <≡>T represent disablement, causation, and state change primitives, respectively. Right: Six example BABEL sentences generated from the CD structure on the left grouped into three paraphrase pairs. Each paraphrase sentence pair has a Jaccard distance greater than 0.94, indicating a large surface difference.

((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."

ing to the syntactic context in which the word sense can be generated and information about how it is generated grammatically. In cases where the word sense may have subordinate clauses or phrases, the concexicon contains corresponding grammar symbols, as well as pointers to substructures of the CD structure with 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 node of the head word sense is furnished with links to nodes generated by recursive calls. Once semantic network generation is complete, 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 (Goldman, 1975).

We ran BABEL in its AND-OR paraphrasing mode, which generates every English realization of a CD structure. After modifications and additions to the original conceptual lexicon and grammar, we generated 432 sentences, from which all possible matchings gave us 93,096 sentence pairs for a paraphrase corpus. Table 2 shows the Conceptual Dependency structure used and examples of the BABEL sentences.

3.3 Paraphrase Recognition Systems and Models

Our experimental comparison also featured both existing paraphrase recognition systems—to represent the alignment paradigm—and a study with human subjects—to represent the deep understanding paradigm. This subsection describes these analyses of the paraphrase recognition task.

3.3.1 BERT

As our exemplar of an alignment-based paraphrase classifier system, we used Devlin et al.'s (2019) BERT (Bidirectional Encoder Representations from Transformers). BERT has been called a "general purpose architecture for natural language understanding" and the system has achieved perfor-

mance scores that are substantially better than its predecessors on a number of tasks that involve natural language processing and on a variety of data sets.

BERT's transformer sequence-to-sequence model architecture allows for efficient pre-training of language models over large unlabeled language corpora through parallelization, followed by a second "fine-tuning" training phase on a smaller labeled data set for a specific task. BERT's distinctive advances for NLP were its elimination of unidirectionality constraints, its greater ability for learning long-range relationships in text, and its presentation of a unified architecture capable of multiple language-processing tasks, including classification and generation. For our studies we used a version called $\text{BERT}_{\text{BASE}}$ that had been fine tuned to perform the MRPC classification task and $\text{BERT}_{\text{BASE}}$ had 110 million total parameters.[2] We ran BERT on the paraphrase recognition task for all 93,096 paraphrase pairs in the BABEL corpus.

3.3.2 Studies with Human Subjects

We chose humans as our exemplars of paraphrase recognizers for the deep understanding paradigm and carried out an experiment with human subjects who performed paraphrase recognition. Some 208 participants took part in the study through an Amazon Mechanical Turk human intelligence task (HIT). These subjects were "Masters" Turk workers who had a 90% approval rate and at least 1,000 previously approved HITs. Participants in the experiment filled out a brief survey that presented the respondents with pairs of English sentences and asked them questions about whether sentences in the pairs had similar meaning.

The survey presented each participant with a BABEL sentence pair and an MRPC sentence pair. Participants rated each 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 they were not simply answering randomly. The attention check items were created by sampling sentences from entirely different data sets in the GLUE benchmark and manually matching them so that the sentences in the pair were obviously on different topics. Submissions from participants who did not classify these items as 0 on the STS scale were rejected and discarded. We counterbalanced the ordering of the sentence pairs to eliminate ordering effects in participants' answers.

The 208 BABEL sentence pairs 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 that had the maximum Jaccard distance of 0.957. In the third, we chose the 40 sentence pairs which had BLEU scores ≤ 0.4. We chose these groups to assure that we had considerable 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 that were marked as paraphrases in the original MRPC annotation.

2. A larger model, $\text{BERT}_{\text{LARGE}}$, was available but tests showed that it only produced a tiny increase in performance over $\text{BERT}_{\text{BASE}}$. We used the Huggingface Transformers system (`https://github.com/huggingface/transformers`), which has an easy-to-use interface to a BERT instance tuned for the MRPC classification task.

4. Results on Alignment and Deep Understanding

Our main experiment tested whether an alignment-based system could detect paraphrases as easily as a deep understanding system when the two sentences in a paraphrase pair are not aligned lexically and syntactically. To test our hypothesis, we used a system from the deep understanding paradigm (BABEL) to generate paraphrase pairs with a range of lexical and syntactic difference, and analyzed the behavior of a learned neural network trained on an alignment-based paraphrase corpus (MRPC) as it performed the paraphrase detection task. In addition, we analyzed our human subjects' data to confirm that our generated paraphrase pairs really were paraphrases and to compare them to paraphrases in alignment-created data sets. We also calculated measures of linguistic difference between paraphrase pairs in order to correlate them with alignment model performance. This section describes these results and analyses.

4.1 Sentence Length and Distance Measures

To compare alignment-based MRPC paraphrases and understanding-based BABEL paraphrases at a lexical and syntactic surface level, we computed sentence lengths and sentence pair distances for each. Before comparing sentence pairs based on their distances, we checked to see if they were comparable in terms of their lengths. We found that the number of word tokens in the MRPC sentences ranged from 5 to 31 tokens with a mean of 18.9. The BABEL sentences were slightly shorter; they had a mean of 12.3 tokens per sentence and ranged from 7 tokens to 18 tokens. However, due to the variance in sentence lengths in both data sets, we did not believe that the average difference in length would be a factor in our results.

We also 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 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 5,801 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 zero, indicating that they used exactly the same words. Table 1 shows an example of an MRPC sentence pair with zero Jaccard distance. This illustrates how alignment methods for creating paraphrases often result in sentence pairs that are nearly or completely the same lexically, with only minor syntactic differences that rearrange large phrases in a sentence.

The BABEL sentences had a Jaccard distance mean of 0.7 and a median of 0.72, ranging from 0 to 0.957. 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, finding the Jaccard distance was significantly larger statistically for the BABEL sentences ($U \simeq 2.0 \times 10^8$,

Table 3. Semantic textual similarity scores on BABEL and MRPC sentence pairs from our 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 %	–	–	–	–

$p < .001$). This indicates that, despite a large percentage of the MRPC sentences being classified as non-paraphrases, the BABEL sentence pairs, all of which are paraphrases of each other, exhibited a greater degree of lexical difference overall. This is an important factor when considering how the training data for alignment-based paraphrase classifiers affects their performance on deep understanding-based paraphrases, as discussed in Section 5.

4.2 Human Classifications and BERT Classifications

In our main experiment we observed both deep understanders (humans) and alignment-based systems (BERT) performing paraphrase detection tasks on paraphrase pairs of both types. First, we conducted studies with human subjects to confirm that our generated paraphrase pairs really were paraphrases and to compare them to paraphrases in alignment-based data sets.

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 time with a mean STS scale score of 4.9 (Table 3) . Participants only rated the BABEL sentences less than 4 in two cases. In contrast, human evaluators found the 208 MRPC sentences 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 one-fifth as "roughly equivalent", with a mean STS score of 4.05, and a median of 4. As a result, we can conclude that the BABEL sentences pairs truly are paraphrases or they can be considered paraphrases at least as much as the MRPC sentences.

In contrast to the human evaluators, when we ran a BERT model trained on the MRPC data set on the BABEL sentence pairs, it classified only 82.1% of them as paraphrases. Given that 100% of the BABEL sentence pairs were constructed to be paraphrases and confirmed to be paraphrases by human evaluators (either "completely equivalent" or "mostly equivalent"), this could be considered an accuracy of 82.1% on the task. This is in comparison to an accuracy of 88.2% that BERT is reported to achieve on the MRPC paraphrases. Although this appears to be a reasonable level, a deeper analysis (in the following section) shows that BERT is highly dependent on surface similarity of the sentence pairs in making its judgments, and is particularly poor at recognizing two sentences as paraphrases when they are lexically and syntactically dissimilar.

4.3 Correlation Measures

We performed further analyses in connection to our hypothesis that a surface alignment system is unable to detect paraphrases when the sentences were not aligned lexically and syntactically. To better understand the effect of surface difference on how humans and BERT determined whether two sentences were paraphrases, we calculated the correlations between BERT's and humans' classifications of both BABEL and MRPC paraphrases, as well as distance measures between the pair of sentences they were classifying. We also calculated correlation tests between distance measures of sentences in paraphrase pairs in the MRPC training data and the original annotations to determine how much BERT's paraphrase detection behavior could be explained by training and tuning.

4.3.1 Correlations between Distance Measures and BERT Classifications

To determine the degree to which BERT detects paraphrases by comparing whether the two sentences in the pair use the same words, we calculated the percentage of BABEL sentences that it 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 classified 100% of the pairs as paraphrases. However, the percentage on the BABEL sentences declined to 80.1% when only considering sentence pairs with edit distance greater than eight. In particular, for 68 BABEL pairs with 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 these same sentence pairs was 4.94. They found them to be "completely equivalent" 94.1% of the time and "mostly equivalent" 5.9% of the time.

To find out 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 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 pairs as a function of BLEU distances (Figure 1). 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 1). The trend in BERT classifications indicates that lexical similarity is a major cue in its behavior, while the human annotators use a meaning representation of the sentences to make determinations.

We also calculated whether BERT's classifications were statistically correlated to the Jaccard distance or edit distance between the sentences in the pairs. To do this, we split the BABEL paraphrase pairs into two groups: one for pairs of BABEL sentences that BERT classified as paraphrases and another for pairs that BERT classified as non-paraphrases. We performed a Mann-Whitney U test, comparing the mean of the Jaccard distances of the sentence pairs in one group to the mean of the Jaccard distances of the other as interval dependent variables. We found a statistically significant difference ($U \simeq 1.95 \times 10^9$, $p < .001$). We performed the same test for lexical edit distance and again found a significant difference ($U \simeq 2.24 \times 10^9$, $p < .001$). Because the Jaccard distance measures whether sentences in the pair use similar words, these results indicate that BERT could simply be examining the words in the sentences instead of their meanings. The difference between BERT and humans on BABEL classifications in Figure 1 was statistically significant.

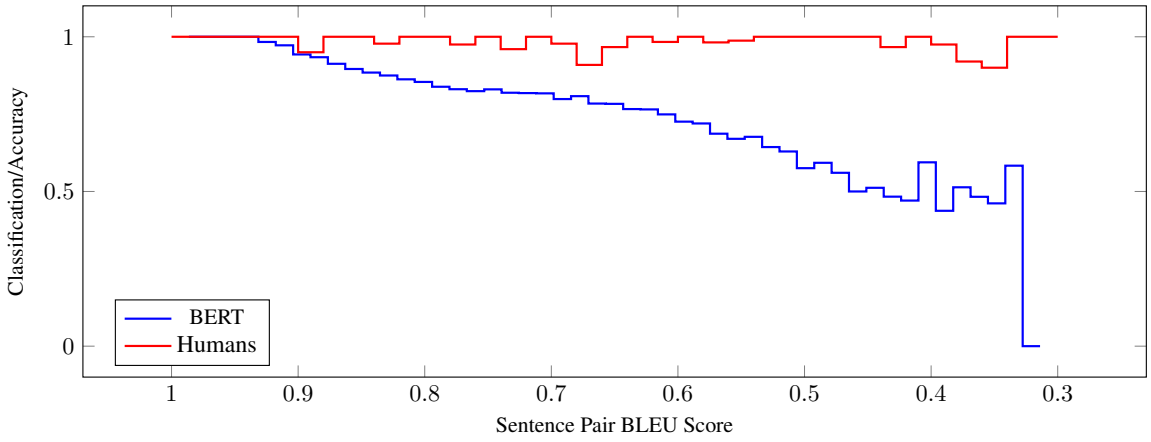

Figure 1. Accuracy of BERT and human annotators for BABEL sentence pairs as a function of BLEU score for the pair. Smaller values of BLEU indicate less lexical overlap or greater lexical "distance". Human classifications are STS classifications (0-5) that have been scaled to the 0–1 classification range 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 equal-sized over the range of BLEU scores. The trend indicates that lexical similarity is a major cue for BERT.

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 its training data for paraphrase recognition, the MRPC, we again performed a Mann-Whitney U test, but this time we compared the Jaccard distances for the 67% of the original 5,801 MRPC pairs that were annotated as paraphrases against the 33% that were not. We found that both Jaccard distances and edit distances were statistically larger for the non-paraphrase pairs in MRPC: 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 it was 0.57 for non-paraphrase pairs. If BERT classified sentence pairs as paraphrases by simply examining whether the sentences used the same words, the reason may be that the same patterns are present in the MRPC training data.

The 1,500 sentences in the Semantic Textual Similarity data set (Agirre et al., 2012) are MRPC sentences annotated with "gold standard" STS scores—for each sentence pair, the annotation is an average of five scores given by crowdworkers. We calculated a Spearman's ρ correlation between Jaccard distance and the gold standard score in the STS data set, 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 STS score ($r_s = -0.115$, $p < 0.001$). Higher STS scores, which indicate that crowdworkers saw the sentence pairs as conveying the same or similar meaning, were correlated with lower Jaccard distances, which indicate that the sentences in the pair used the same words. This is further evidence that BERT's bias toward viewing lexically similar sentences as paraphrases may be due to the MRPC data on which it was trained.

4.3.3 Lack of Correlations between Distance and Human Classifications

We saw in Figure 1 that human raters appeared 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 support for this assumption, we ran a Spearman's ρ test for a correlation between human ratings of the BABEL sentences and Jaccard distance and found no correlation ($r_s = 0.03$, $p = 0.6$). Similar tests for correlations between human ratings and edit distance and for correlations between human ratings and BLEU score produced similar results. We conclude that human raters' skill at detecting that two sentences are paraphrases is independent of whether the sentences in the pair are lexically or syntactically similar, indicating that they are manipulating non-linguistic representations when understanding sentences for paraphrase classification.

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 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 like BABEL that more accurately resemble human language behavior.

We also argue for our claims based on examination of how paraphrasing is 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, humans might match at a sentence-wide level and store separate relation entries for the 93K BABEL sentence pairs or any other possible pair that could be encountered in the task. However, this option seems implausible given the uncountable character 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 its inefficiency, since an equivalence between n expressions would require $O(n^2)$ relation entries. Establishing equivalence through chains of relations cuts down on this expense. But with certain linguistic expressions serving as a representative of other linguistic expressions in a chain, this solution begins to resemble the separate, non-linguistic meaning representation of deep understanding.

On the other hand, with deep understanding, if language inputs are transformed into 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 by noting how BABEL's decoupling of linguistic knowledge from non-linguistic CD conceptual representations facilitates the generation of so many paraphrases combinatorically from the sparse input of a single CD structure.

But it is also important to consider the language understanding tasks that researchers are focused on when training and testing models like BERT, and whether the training data sets express the full range of human language. We found that not even a majority of the MRPC sentences originally

annotated as paraphrases were considered "completely equivalent" on the STS scale by our participants. This seems to be a major drawback of the alignment method for creating paraphrase data sets.

In the literature on textual similarity and paraphrase detection, the main clues pointing to the challenges of the alignment account of human language involve evolution of the classification systems themselves, which retreat from expressions of meaning equivalence to embrace vague gradations of "semantic relatedness" and "meaning overlap". Dolan and 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 is that the original method for constructing MRPC attempts to align sentences from a cluster of news stories on a particular event. Many news stories are only "minor edits" of an original AP or Reuters story. Also, different 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 seldom find pairs that express nearly exactly the same idea and that vary significantly lexically and syntactically but not in the details that they express.

In this study, we only subjected BERT to sentence pairs that were paraphrases generated from the non-linguistic representation. We did not construct non-paraphrases with a range of syntactic and lexical similarity and difference in their constituent sentences and subject the model to them. This may have created a more balanced picture of the system's behavior. Also, stemming, lemmatization, removal of stop words, and other measures may have let us achieve more precise measurements and results linguistically. We used edit distance as a measure of sentence similarity because it contributed heavily to extraction of the MRPC data set and because it is a measure that incorporates syntactic differences. But edit distance does not distinguish between lexical and syntactic differences. A purely syntactic measure (e.g., Tree Kernels, Collins & Duffy, 2002) would complement lexical ones that are based solely on bags of words, such as BLEU and Jaccard distance.

6. Related Work

An abundance of prior work exists on systems and methods for recognizing, generating, and extracting paraphrases, as well as the related task of textual entailment. A survey by Androutsopoulos and Malakasiotis (2010) covers paraphrasing systems and textual entailment, including ones that are handcrafted. Madnani and Dorr (2010) also survey data-driven approaches to alignment-based paraphrase generation. While performing a manual examination of MRPC, Bhagat and Hovy (2013) attempt to avoid conflict between narrow and broad notions of paraphrasing to define "quasi-paraphrases" that convey "approximately the same meaning".

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 been trained to identify paraphrases and predict sentence similarities, as well as to generate paraphrases (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 well these systems capture syntax and surface-level linguistic forms, not how they represent meaning.

The literature also includes examples of alignment-style extraction of paraphrase sentence pairs. Lin and Pantel (2001) confront alignment from an information extraction perspective, while Barzilay and McKeown (2001) align sentences in multiple English translations of classic literary works. Barzilay and Lee (2003) calculate an alignment form called "lattices"—graph-based representations of structural similarity in sentences in a corpus of news articles—and pair sentences as paraphrases if they share entities as arguments. Pang et al. (2003) extract synonyms and phrasal paraphrases from multiple translations of news articles as finite-state automata and report a human evaluation, but only for synonyms and short phrasal paraphrases represented by these structures.

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. Spärck Jones and Tait (1984) use a linguistically-motivated generator to provide a systematic range of expressions of indexing concepts for document retrieval, while McKeown (1979) draws on hand-crafted syntactic expertise to generate paraphrases of natural language database queries to aid infrequent users. However, these systems are limited to shallow syntactic representations of meaning, whereas our work continues an alternative thread of research on paraphrase generation based on meaning representations.

7. Conclusion

In this paper, we supported claims about the characteristics of meaning representation systems that are required to perform natural language understanding and generation. We argued that symbolic structures which are distinct 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 in support of these claims.

We analyzed extensively 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 and that exhibited a wide range of syntactic and lexical difference in their constituent sentences. We found that it deviated significantly from human classifications, particularly on the sentence pairs that differed linguistically. These results imply that either the neural models or the data sets used to trained them are missing fundamental components of meaning representation.

Work on adversarial examples for neural networks, rather than convincing researchers of the drawbacks of research that focuses on shallow performance metrics, has recently turned to methods for generating more training data that ostensibly make models more robust. We resist the urge to turn sentence pairs created by BABEL into data for tuning BERT in paraphrase detection tasks. Although our study shows that BERT's behavior may only reflect weaknesses in the MRPC, we think a more important avenue of future investigation is to train a model that encodes the non-linguistic aspects of meaning present in BABEL. This would require, for example, training a system corresponding to BABEL's semantic network system, which builds linguistic surface realizations via a mapping from the non-linguistic meaning representations. A major challenge of this would be to acquire data sets large enough to train a neural network.

This paper's accomplishments were only possible because we expended the effort to build a symbolic system by hand as an expression of a meaning representation theory, rather than building a large data set. While some AI traditions have grown to fear these kinds of systems because of the "knowledge bottleneck" and impractical prospects for scaling them into a complete human-level intelligence, we are not discouraged by these prospects. We can use hand-built systems as simulation tools to test theories at a small scale, which we can use later to enhance or complement data-driven learning approaches. Systems like BERT on their own do not appear to lead to a better understanding of human language processing. Instead, we intend to investigate other data sets and models for natural language interpretation through the kind of study presented here, while also exploring improvements to the non-linguistic representation of meaning.

Acknowledgments

We thank Neil Goldman and Bruce Baumgart for providing the original BABEL code through the SAILDART archive project (`https://www.saildart.org/`) of the Stanford Artificial Intelligence Laboratory.

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, LA: AAAI Press.

Collins, M., & Duffy, N. (2002). Convolution kernels for natural language. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15*, 625–632. San Diego, CA: The NIPS Foundation.

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, MN: 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). Lisbon, Portugal.

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

Iyyer, M., Wieting, J., Gimpel, K., & Zettlemoyer, L. (2018). Adversarial example generation with syntactically controlled paraphrase networks. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1875–1885). New Orleans, LA: Association for Computational Linguistics.

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). Santa Fe, NM: COLING.

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, CA: 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 the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 2923–2934). Osaka, Japan: COLING.

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., Pennington, J., Ng, A. Y., & Manning, C. D. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Proceedings of the 24th International Conference on Neural Information Processing Systems* (p. 801–809). Red Hook, NY: Curran Associates.

Spärck 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.

Woods, W. A. (1970). Transition network grammars for natural language analysis. *Communications of the ACM*, *13*, 591–606.

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). Denver, CO: Association for Computational Linguistics.