Mutual Implication as a Measure of Textual Equivalence

Animesh Nighojkar, John Licato
Advancing Machine and Human Reasoning (AMHR) Lab
Department of Computer Science and Engineering
University of South Florida, Tampa, FL, USA
{anighojkar, licato}@usf.edu

Abstract

Semantic Textual Similarity (STS) and paraphrase detection are two NLP tasks that have a high focus on the meaning of sentences, and current research in both relies heavily on comparing fragments of text. Little to no work has been done in studying inference-centric approaches to solve these tasks. We study the relation between existing work and what we call *mutual implication* (MI), a binary relationship between two sentences that holds when they textually entail each other. MI thus shifts the focus of STS and paraphrase detection to understanding the meaning of a sentence in terms of its inferential properties. We study the comparison between MI, paraphrasing, and STS work. We then argue that MI should be considered a complementary evaluation metric for advancing work in areas as diverse as machine translation, natural language inference, etc. Finally, we study the limitations of MI and discuss possibilities for overcoming them.

Introduction

What does it mean when we say that two sentences mean the same? According to the *meaning-as-use* view (Wittgenstein 1953), the meaning of a word in a language game, or of a symbol in a representational system, comes from how it is (or can be) used. In contemporary natural language processing (NLP), approaches to computing word meaning and two well-studied tasks have been overwhelmingly influenced, albeit indirectly, by this view of meaning: *semantic textual similarity* (STS), which attempts to assign a numerical measure of similarity between two text segments; and *paraphrase detection*, which attempts to identify whether two statements are paraphrases of each other.

Much work in both STS and paraphrase detection relies on approaches that rely on overlaps between words, n-grams, or other subsets of the texts being compared. This paradigm has been quite successful, particularly since the introduction of transformer-based architectures and their contextual embeddings (Vaswani et al. 2017; Devlin et al. 2018; Zhang et al. 2019). However, understanding the ways in which sentences are or can be used requires that we consider the holistic meanings of sentences, rather than just the separate meanings of their constituent words; e.g., the aspects of sentence meaning that are inference-centric. This is the approach recommended by *inferential role semantics*, or *inferentialism* (Boghossian 1994; Peregrin 2006), whereby the meaning of a statement s is grounded in its inferential properties: what one can infer from s, and from what s can be inferred.

In this paper, we initiate an exploration into the degree to which current work in NLP captures inferential role semantics. In particular, we will study what we call *mutual implication* (MI), a binary relationship between natural language sentences that holds when each sentence textually entails the other. For this paper, we use an estimate of textual entailment: a RoBERTa model trained on a combination of natural language inference (NLI) corpora. For convenience, in much of this paper when we say that two sentences are MI (or that they are "mutually implicative"), we mean that they were determined to textually entail each other using this RoBERTa model.

Mutual implication is worth studying for many reasons. It focuses on inferential relationships between sentences, which are holistic properties of the sentences and how they relate to each other and to the background knowledge. This is in contrast to many STS benchmark datasets, which tend to reward reliance on surface-level feature, word, or n-gram comparisons. One might expect that in the limit, STS scores reach toward MI; i.e., a sentence pair with an extremely high STS score should be extremely likely to be MI, and vice-versa. We show in this paper that this expectation is not straightforwardly met by current STS models.

MI is also very closely related to the concept of paraphrase, and some authors have defined paraphrase in a way that much resembles what we call MI (Marsi, Krahmer, and Bosma 2007; Androutsopoulos and Malakasiotis 2010). However, paraphrase has been defined in non-inferential terms as well (Bhagat and Hovy 2013), so for clarity, we will classify MI as a *type* of paraphrase. Compared to other definitions of paraphrasing (e.g., those based on STS), MI has the advantage of providing a sharp, non-arbitrary boundary for detecting *non*-paraphrases: if textual entailment fails to hold in either direction, then the sentences are not MI. Because this boundary is sharp and lends itself to explainability (e.g., demonstrating why a textual entailment fails to hold can be explained with the use of counterexamples), MI

Corpus	$s_1 \rightarrow s_2$	$s_2 \rightarrow s_1$	$s_1 \leftrightarrow s_2$
ParaNMT	75.82	60.49	50.21
PPNMT	78.44	77.27	68.15
MSRP (paraphrases)	39.30	41.74	17.98
MSRP (non-paraphrases)	10.73	12.09	0.68

Table 1: Percentage of sentence pairs from paraphrase datasets that entail each other, along with the non-paraphrase subset of MSRP for comparison

as a type of paraphrase may have downstream applications in automatic tutoring systems, automatic grading of essays, and plagiarism detection, to name a few.

Contributions of this work The novel contributions of this paper are as follows. We:

- measure the degree to which current paraphrase datasets capture MI,
- measure the degree to which current SOTA STS datasets and models capture MI,
- show that MI can serve as a supplemental measure of the quality of machine translation systems,
- present an updated version of the ParaNMT dataset (PPNMT) that we will release publicly, and
- synthesize the evidence we present to argue that MI should be a parallel goal of STS, paraphrase, and machine translation work.

Related Work

Paraphrasing involves expressing the same information in multiple ways (Pang, Knight, and Marcu 2003) to achieve varying levels of fluency (Iordanskaja, Kittredge, and Polguère 1991), clarity, and summarization (McKeown et al. 2002). Whereas textual entailment is more closely tied to reasoning and inference, paraphrasing tends to be broader, encompassing not only entailment but substitutability, preservation of information, etc. (Androutsopoulos and Malakasiotis 2010; Bhagat and Hovy 2013; Mingers 1995). Bhagat and Hovy (2013) refer to this form of paraphrase as *quasi-paraphrases*, and we suspect that the sense of paraphrase currently dominant in NLP research is closer to this. For example, consider the sentences "It's easy if you book one of our guided tours," and "It will be better if you book one of our guided tours." Although these are each considered paraphrases of the other (according to classifiers we will discuss later), neither textually entails the other.

Many paraphrase datasets exist: two of the most prominent including ParaNMT and MSRP. ParaNMT (Wieting and Gimpel 2018) was created by machine translating the Czech side of a human-translated Czech-English parallel corpus (Bojar et al. 2016) using a neural machine translation (NMT) system (Sennrich et al. 2017), such that the human and machine translations are paraphrases of one another. When referring to sentence pairs in ParaNMT data set, we will say that s_1 and s_2 are the human- and machine-translated sentences, respectively. Microsoft Research Paraphrase Corpus (MSRP) (Dolan and Brockett 2005) contains 5801 sentence pairs, each with a binary number showing whether humans considered those two sentences paraphrases. Unlike ParaNMT, MSRP contains explicitly non-paraphrase sentence pairs too.

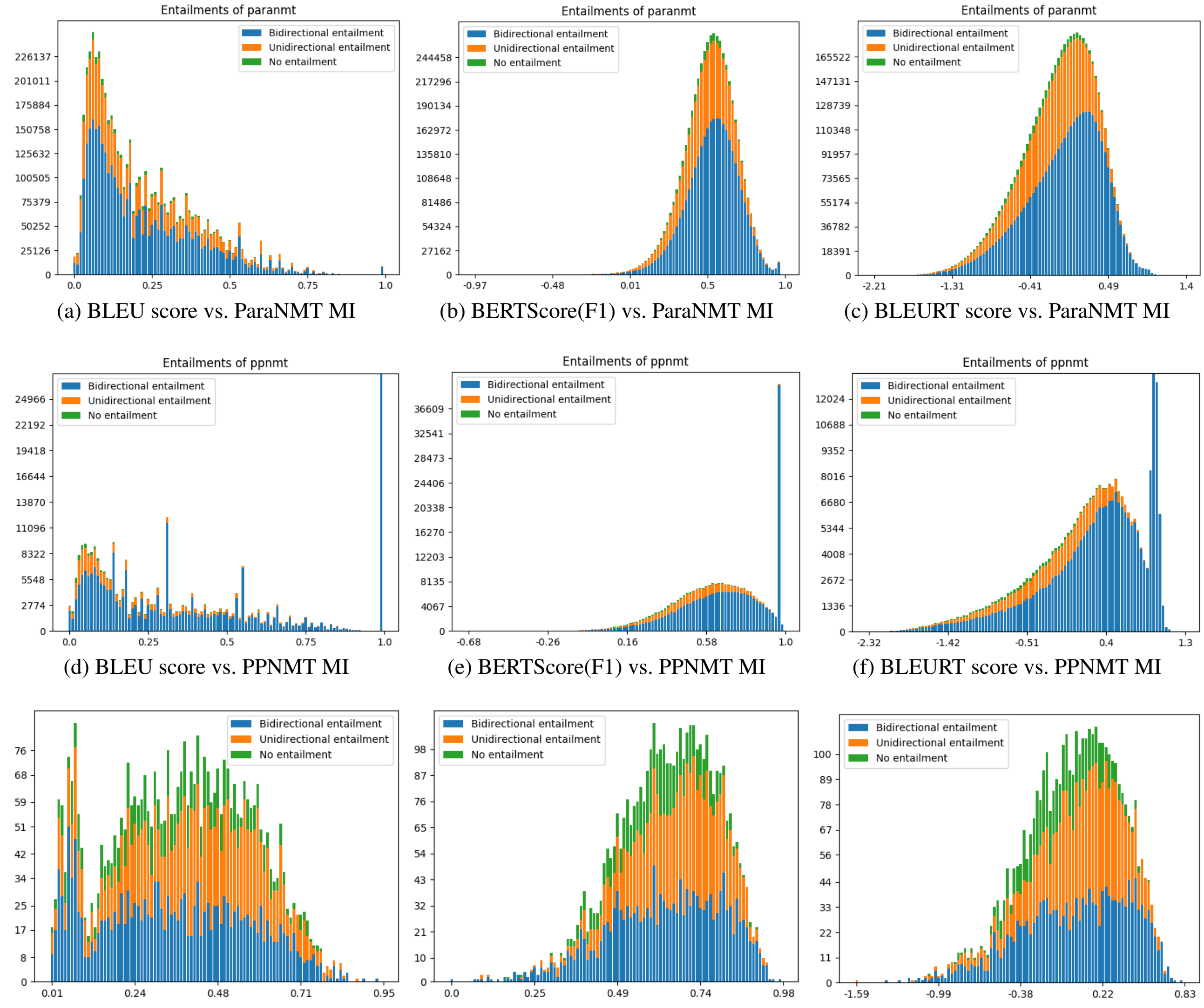
Semantic textual similarity (STS) (Agirre et al. 2012) measures the degree of semantic similarity between two given sentences. STS is clearly related to both paraphrasing and textual entailment, and is applicable to areas like machine translation (Cer et al. 2017) and question answering (Lan and Xu 2018) due to its utility in detecting minor semantic differences. Common approaches to calculating STS include word error rate (Levenshtein 1966; Panja and Naskar 2018; Stanchev, Wang, and Ney 2019) and n-gram matching (e.g., BLEU (Papineni et al. 2002)).

More recent approaches such as BERTScore (Zhang et al. 2019) draw on contextual word embeddings derived from BERT (Devlin et al. 2018), thus allowing them to recognize not just semantically similar words but similar phrases and synonyms. As a result, BERTScore produces STS scores closely matching human judgments of similarity. Ultimately, however, BERTScore still performs token-token matching to compute the precision and recall and thus is sensitive to sentence structure. BLEURT (Sellam, Das, and Parikh 2020) is a text generation metric which builds on BERT's contextual word representations. BLEURT is "warmed-up" using millions of synthetic sentence pairs twice (on Language Modeling (Devlin et al. 2018) and then on Natural Language Generation evaluation) before it is fine-tuned on human ratings, thus getting a better performance than other STS metrics with a high correlation with human ratings.

Natural language inference (NLI) (Bowman et al. 2015; Williams, Nangia, and Bowman 2018), sometimes referred to as recognizing textual entailment (RTE), is the task of determining whether a hypothesis h can be inferred given a premise p. E.g., given $s_1 =$ "Two black cars start racing in front of an audience." and some s_2, the possible relationships are:

1. *Entailment*: $s_1 \rightarrow s_2$. Based on any or all information in s_1, s_2 can be said to be true. $s_2 =$ "Two cars are racing."
2. *Contradiction*: Based on any or all information in s_1, s_2 can be said to be false. $s_2 =$ "A man is driving down a lonely road."
3. *Neutral*: Based on all information in s_1, s_2 can be either true or false (insufficient information). $s_2 =$ "Two men are racing in black cars."

In contrast to STS, the focus of NLI is primarily inferential. NLI Datasets, like SNLI (Bowman et al. 2015), MultiNLI (Williams, Nangia, and Bowman 2018), ANLI (Nie et al. 2020), etc., meant to capture NLI and train language models have received much attention recently. SNLI consists of 570k sentence pairs, all of which were written as well as annotated by humans via Amazon mTurk. MultiNLI consists of 433k sentence pairs modeled on SNLI but with a focus on having sentences from multiple genres of spoken and written text. It has the same format as SNLI and the creators of



(a) BLEU score vs. ParaNMT MI

(b) BERTScore(F1) vs. ParaNMT MI

(c) BLEURT score vs. ParaNMT MI

(d) BLEU score vs. PPNMT MI

(e) BERTScore(F1) vs. PPNMT MI

(f) BLEURT score vs. PPNMT MI

(g) BLEU score vs. MSRP MI

(h) BERTScore(F1) vs. MSRP MI

(i) BLEURT score vs. MSRP MI

Figure 1: Number of mutual implications (bidirectional entailment), single-direction entailment, and non-entailment of the ParaNMT, PPNMT, and MSRP datasets based on BLEU score, BERTScore (F1 pictured only, as precision and recall produced similar graphs), and BLEURT score. All figures divide the range of observed scores into 100 bins.

MultiNLI suggest using both the corpora together as a single large corpus. ANLI (Adversarial NLI) is a dataset comprising of around 169k (19k + 47k + 103k) sentence pairs from three rounds of adversarial, iterative, ”human-and-model-in-the-loop” procedure. The humans’ role in this training process is to devise examples which fool the model, which are then added to the training set to train a stronger model. This makes ANLI one of the hardest NLI datasets at present. We use a pre-trained RoBERTa-large (Liu et al. 2019) model launched by (Nie et al. 2020) trained on SNLI, MultiNLI, FEVER-NLI (Nie, Chen, and Bansal 2019), and ANLI (all 3 rounds). Given two sentences (s_1, s_2), we use this RoBERTa model to predict entailment in both directions. If entailment is predicted in only one direction, we say that the sentence pair has *unidirectional entailment*; if predicted in both, we say that the sentence pair is MI.

Experiments

Are paraphrases mutually implicative?

As stated earlier, MI can be considered a type of paraphrase relationship, but the concept of paraphrase in general is somewhat inconsistently defined. We therefore expect that contemporary paraphrase datasets may be a mix of MI and non-MI sentence pairs, and disentangling these two subsets from each other may help us understand more about the differences between the concepts of MI and paraphrase in general. We do this by first analyzing the ParaNMT dataset (`ParaNMT-5M-processed`, specifically) and the

BLEURT	Sentence 1	Sentence 2
-0.801	But Odette is the first to form over the Caribbean Sea in December, the Center said.	It is the first named storm to develop in the Caribbean in December.
-0.767	Sens. John Kerry and Bob Graham declined invitations to speak.	The no-shows were Sens. John Kerry of Massachusetts and Bob Graham of Florida.
-0.617	The judge ordered the unsealing yesterday at the request of several news agencies, including The Seattle Times, The Associated Press and the Seattle Post-Intelligencer.	The depositions were made public yesterday at the request of the P-I, The Seattle Times and The Associated Press.

Table 2: Selected MSRP sentence pairs identified as MI but with low BLEURT scores

BLEURT	Sentence 1	Sentence 2
0.738	NBC probably will end the season as the second most popular network behind CBS, although it's first among the key 18-to- 49-year-old demographic.	NBC will probably end the season as the second most-popular network behind CBS, which is first among the key 18-to-49-year-old demographic.
0.591	The 30-year bond US30YT=RR dipped 14/32 for a yield of 4.26 percent from 4.23 percent.	The 30-year bond US30YT=RR lost 16/32, taking its yield to 4.20 percent from 4.18 percent.
0.562	Advancing issues outnumbered decliners nearly 2 to 1 on the New York Stock Exchange.	Declining issues outnumbered advancers slightly more than 3 to 1 on the New York Stock Exchange.

Table 3: Selected MSRP sentence pairs identified as not MI but with high BLEURT scores

Microsoft Research Paraphrase Corpus (MSRP). Using the previously described RoBERTa model for recognizing entailment, and given sentence pairs (s_1, s_2), we determine for what percentage of the sentence pairs (1) s_1 entails s_2, (2) s_2 entails s_1, and (3) both. Table 1 lists the results.

A few results from Table 1 are interesting to note. ParaNMT, which purports to consist entirely of paraphrase pairs, only passes the MI test 50.21% of the time, performing as good as chance. The subset of MSRP consisting only of those sentence pairs which were labeled as paraphrases does much worse: only 17.98% were identified as MI. These results are consistent with our view that MI is one sense of paraphrase which datasets like ParaNMT and MSRP do not capture effectively. No doubt, some of these results are due to the limitations of our MI classifier and the data it was trained on, but the results presented here can be considered a baseline—both for NLI models and paraphrase generators—against which future work can compare.

MI as an NMT Evaluation Metric

An interesting asymmetry can be observed in Table 1: the ratio of paraphrase pairs for which s_1 entails s_2 is considerably higher than it is for s_2 entailing s_1. We suspect this is because of how ParaNMT was constructed—s_1 is the translation from Czech to English written by humans and s_2 is the translation from Czech performed by NMT (Neural Machine Translation). The fact that the textual entailments are asymmetric suggests that there is some sort of information loss in the NMT process; and therefore, *the degree to which NMT minimizes this asymmetry, and maximizes MI, may constitute a new way of assessing machine translations*.

Given the rapid progress in the field in the past few years, ParaNMT's reliance on translation algorithms that are a few years old may be to blame for the asymmetry in Table 1. We hypothesize that a dataset of sentence pairs using updated machine translation models will have a smaller entailment asymmetry and higher MI ratio than that observed with ParaNMT. To test this idea, we use the google-translate library [1] to translate the Czech side of the CzEng 2.0 (Kocmi, Popel, and Bojar 2020) corpus which has filtered CzEng 1.6 and six additional resources. Due to resource limitations, we translate only the test corpus of CzEng 2.0 (roughly 300K pairs), and call this data set *Para-ParaNMT* (PPNMT for short, the prefix *para-* reflecting its similarity to, and conceptual derivation from, ParaNMT).

The results are listed in the second row of Table 1. As expected, the asymmetry gap for PPNMT is greatly reduced (1.17% as compared to 15.33% with ParaNMT), and the percentage of sentence pairs that are MI is increased to 68.15% from 50.21%, allowing us to conclude that the asymmetry detected by calculating MI does indeed seem to reflect the improvement in machine translation, supporting the idea that it can be used as a measure of translation quality.

Does STS capture entailment?

To further visualize the relation between STS scores and MI, we calculate the BLEU, BERTScore, and BLEURT scores for sentence pairs in the ParaNMT, PPNMT, and MSRP datasets. Dividing the range of observed scores into 100 bins, we then calculate the number of sentence pairs in each bin which are MI, have unidirectional entailment, and have no entailment. Figure 1 contains the resulting stacked bar charts. Two differences immediately stand out: (1) The ratio of MI to non-MI pairs is much higher for PPNMT than with the other datasets. This is expected with MSRP, as it is the only one of the three datasets to intentionally contain non-paraphrase pairs. (2) The degree to which each STS score predicts MI can be seen by the skew of the inner blue curve towards the right. E.g., this skew is more pronounced in Figure 1c than in 1b. The spikes to the right of the PPNMT figures correspond to sentence pairs which were almost exactly identical; the lack of these in ParaNMT further demonstrates how much NMT methods have improved in the past

[1] www.pypi.org/project/googletrans/

few years.

Although a vast majority of sentence pairs in all three datasets are entailments in at least one direction, STS metrics seem to fail at capturing this information: the expectation would be that non-entailment, unidirectional entailment, and MI sentence pairs should dominate the lower, middle, and higher score ranges, respectively—an expectation that most prominently fails to manifest in Figures 1g–1i. We interpret this as further evidence of the subtle distinction between inference-centric properties of sentence pairs and the featural comparisons measured by STS measures.

Conclusion and Future Directions

The statements "x and y are mutually implicative" and "x and y are paraphrases" are themselves neither mutually implicative nor paraphrases. Although the former statement does imply the latter, the reverse is not true, and it is important to understand the difference. MI, we have argued, is a type of paraphrase relationship which should be studied and modeled in its own right, as it is inference-centric, lends itself to explainability (through counterexamples to disprove entailment), and has other advantages as well—such as its possible use as a way of assessing the quality of machine translation, which we have demonstrated here.

Although we used RoBERTa, a state-of-the-art language model which we have fine-tuned on the latest NLI corpora (Nie et al. 2020), it is important to remember that it is an imperfect approximator of textual entailment. Limitations of this model can be seen in Tables 2 and 3, which list selected sentence pairs which had BLEURT scores in the lower and upper 5th-percentiles (respectively), but were identified as MI or non-MI (respectively). Manual inspection of most sentence pairs which had high BLEURT scores but were not considered MI (and low BLEURT / considered MI) appear to have been incorrectly classified by our RoBERTa model, and is thus a limitation of the work we present here. We expect that this limitation can be alleviated as research in NLI continues (Raffel et al. 2019; Lan et al. 2019; Yang et al. 2019; Liu et al. 2020; Nie et al. 2020; Wang et al. 2020; Brown et al. 2020), but we argue that this further validates the novel explorations and methods proposed in this paper: This work can be re-visited, and our results used as baselines, to further understand the strengths and limitations of SOTA STS, paraphrase, machine translation, and NLI approaches. Furthermore, our results highlight the need to not only use MI as a method for assessing sentence similarity, but the utility of MI as a way of improving NLI systems as well, e.g. by ensuring that sentence pairs identified as semantically equivalent pass the MI test.

Use of the MI test can be applied to evaluating text generation techniques, the most prominent applications of which are machine translation, speech synthesis, image captioning, etc. However, our argument is not that MI should be used *instead* of STS metrics, but rather that the two approaches are complementary in understanding what makes sentences semantically equivalent. Failures such as those described in Tables 2 and 3 might be reduced if we were to use MI and STS to complement, rather than substitute, each other. Another noteworthy takeaway is that MI is not a model, but a concept which we have shown the practical applications of using a model. Hence we expect further research in NLI to improve MI as well.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award numbers FA9550-17-1-0191 and FA9550-18-1-0052. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

References

Agirre, E.; Cer, D.; Diab, M.; and Gonzalez-Agirre, A. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In * *SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, 385–393.

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

Bhagat, R., and Hovy, E. 2013. What is a paraphrase? *Computational Linguistics* 39(3):463–472.

Boghossian, P. A. 1994. Inferential role semantics and the analytic/synthetic distinction. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition* 73(2/3):109–122.

Bojar, O.; Dušek, O.; Kocmi, T.; Libovický, J.; Novák, M.; Popel, M.; Sudarikov, R.; and Variš, D. 2016. CzEng 1.6: Enlarged Czech-English Parallel Corpus with Processing Tools Dockered. In Sojka, P.; Horák, A.; Kopeček, I.; and Pala, K., eds., *Text, Speech, and Dialogue: 19th International Conference, TSD 2016*, number 9924 in Lecture Notes in Computer Science, 231–238. Cham / Heidelberg / New York / Dordrecht / London: Masaryk University.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners.

Cer, D.; Diab, M.; Agirre, E.; Lopez-Gazpio, I.; and Specia, L. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017).*

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805.

Dolan, W. B., and Brockett, C. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Iordanskaja, L.; Kittredge, R.; and Polguère, A. 1991. *Lexical Selection and Paraphrase in a Meaning-Text Generation Model*. Boston, MA: Springer US. 293–312.

Kocmi, T.; Popel, M.; and Bojar, O. 2020. Announcing czeng 2.0 parallel corpus with over 2 gigawords. *arXiv preprint arXiv:2007.03006*.

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

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations.

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach.

Liu, X.; Cheng, H.; He, P.; Chen, W.; Wang, Y.; Poon, H.; and Gao, J. 2020. Adversarial training for large neural language models.

Marsi, E.; Krahmer, E.; and Bosma, W. 2007. Dependency-based paraphrasing for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, 83–88.

McKeown, K. R.; Barzilay, R.; Evans, D.; Hatzivassiloglou, V.; Klavans, J. L.; Nenkova, A.; Sable, C.; Schiffman, B.; and Sigelman, S. 2002. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, 280–285. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Mingers, J. C. 1995. Information and meaning: foundations for an intersubjective account. *Information Systems Journal* 5(4):285–306.

Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; and Kiela, D. 2020. Adversarial nli: A new benchmark for natural language understanding.

Nie, Y.; Chen, H.; and Bansal, M. 2019. Combining fact extraction and verification with neural semantic matching networks. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Pang, B.; Knight, K.; and Marcu, D. 2003. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. Technical report, CORNELL UNIV ITHACA NY DEPT OF COMPUTER SCIENCE.

Panja, J., and Naskar, S. K. 2018. ITER: Improving translation edit rate through optimizable edit costs. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 746–750. Belgium, Brussels: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Peregrin, J. 2006. Meaning as an inferential role. *Erkenntnis* 64(1):1–35.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Sellam, T.; Das, D.; and Parikh, A. P. 2020. Bleurt: Learning robust metrics for text generation.

Sennrich, R.; Firat, O.; Cho, K.; Birch, A.; Haddow, B.; Hitschler, J.; Junczys-Dowmunt, M.; Läubli, S.; Barone, A. V. M.; Mokry, J.; et al. 2017. Nematus: a toolkit for neural machine translation. *arXiv preprint arXiv:1703.04357*.

Stanchev, P.; Wang, W.; and Ney, H. 2019. Eed: Extended edit distance measure for machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, 514–520.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need.

Wang, B.; Wang, S.; Cheng, Y.; Gan, Z.; Jia, R.; Li, B.; and Liu, J. 2020. Infobert: Improving robustness of language models from an information theoretic perspective.

Wieting, J., and Gimpel, K. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 451–462. Melbourne, Australia: Association for Computational Linguistics.

Williams, A.; Nangia, N.; and Bowman, S. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 1112–1122. Association for Computational Linguistics.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Basil Blackwell.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, 5753–5763.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert.