Pre-Trained Language Models in Semantic Communication

Luiz Fernando Gontijo and Paulo Cardieri

Abstract— **Communication systems traditionally focus on accurately transmitting signals without considering semantic content. This paper introduces semantic communication models using pre-trained language models, T5 and BART, compared to conventional methods like Huffman and Turbo coding. Numerical results demonstrate the semantic models' superiority, especially in low SNR conditions, measured by BLEU and BERTScore metrics. Also, the proposed system only needs fine tuning to obtain good results in environments with severe fading. The results suggests a paradigm shift where decoding semantic meaning, rather than exact message replication, becomes crucial. Such models pave the way for novel communication architectures and emphasize the importance of semantic understanding in communication systems.**

Keywords— **Semantic Communications, Language Models, Signal Processing, Deep Leaning.**

I. Introduction

Communication system models can be defined as the replication of a message at the receiver sent by the transmitter. The model for how we see communication systems was first defined by Shannon in 1948 [1]. This perspective seeks accurate transmission of the message and exact data recovery by the receiver. The search for similarity between messages is proposed at the bit level.

Therefore, conventional communication systems focus on the reliability and efficiency of sending signals and do not consider the semantics present in the message [2]. The transmitter and receiver are treated as agents without intelligence or prior knowledge about the scope of the messages. Messages sent in this way may contain irrelevant and semantically redundant content. Such irrelevant content could be suppressed, if the receiver and transmitter already had prior knowledge about the text or the area of knowledge of the subject sent, resulting in a reduction in the use of communication resources.

Weaver and Shannon defined three fundamental communication problems [3] based on the technical, semantic, and effectiveness telecommunications paradigms. Currently, all communication systems operate with the aim of overcoming the technical problem, and promoting the recovery of symbols by the receiver in the most accurate way possible. As for the semantic problem, the receiving system starts recognizing the symbol sent based on its exact or approximate interpretation. The last problem implies that the receiver of the command sent by the signal conducts itself as expected. Therefore, the semantic-based communication proposal attempts to solve the semantic problem.

Studies on semantic information date back to 1952, carried out by Carnap and Bar-Hillel [4]. However, in recent works, such as [5], attempts have been made to classify alternative semantic representations in different data sources. Therefore, the advancement of AI has directly impacted the development of alternatives for communications networks [8]. The use of deep learning techniques and extensive databases has enabled the achievement of good results in the field of semantic communications, as presented in [9]. In this last work, the advancement of the model using Transformers architecture, initially proposed in [10], is notable.

In the present work, we introduce the use of pre-trained models in semantic-based communication systems. We propose a system with semantic coding and decoding layers obtained by pre-trained BART and T5 models together with channel coding and decoding layers composed of neural networks. The contribution of this work is to show the possibility of using different pre-trained language models to obtain good results in semantic communication models. By using the pre-acquired knowledge capacity of the language models, it was possible to note that the fast training of one epoch was enough for the proposed systems to reach a performance superior to that of traditional coding. We also show that even when the proposed system was trained under non-severe channel conditions (i.e. AWGN channel), its performance overcame the traditional system under severe channel conditions (Rayleigh fading). The results indicate a new research path for the development of semantic communication systems.

II. System Description

A. Problem Description

For the presentation of the proposed system, we assume that the input sentences are represented by $\boldsymbol{s} = [w_1, w_2, w_3, \ldots, w_k]$, where w_l represents the l-th word of the sentence. The encoded signal can be represented by

$$\mathbf{x} = C_\alpha(S_\beta(\boldsymbol{s})), \tag{1}$$

where **x** are complex symbols required for transmission, $S_\beta(\cdot)$ represents the semantic coding network with the parameter set β and $C_\alpha(\cdot)$ is the channel encoding network with the parameter set α.

The symbols **x** are transmitted over a communication channel, which will be disturbed by noise **n** and small-scale fading

Luiz Fernando Gontijo, DECOM-FEEC, UNICAMP, Campinas-SP, e-mail: lz.gontijo@gmail.com; Paulo Cardieri, DECOM-FEEC, UNICAMP, Campinas-SP, e-mail: cardieri@unicamp.br. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nivel Superior – Brasil (CAPES) – Finance Code 001, Grant 88887.673246/2022-00, and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Grant 313213/2021-6.

h. Thus, the channel output is [14]

$$\mathbf{y} = h\mathbf{x} + \mathbf{n}. \tag{2}$$

For the Rayleigh fading channel, the coefficients h follow the distribution $\mathcal{CN}(0, 1)$, while the noise term $\mathbf{n}$ follows $\mathbf{n} \sim \mathcal{CN}(0,\, \sigma^2_{\mathbf{n}})$.

The decoded signal can be represented as

$$\hat{\boldsymbol{s}} = S^{-1}_{\chi}(C^{-1}_{\sigma}(\mathbf{y})), \tag{3}$$

where $\hat{\boldsymbol{s}}$ is the vector with the decoded tokens (smallest units of meaning, such as words and numbers). Furthermore, $S^{-1}_{\chi}(\cdot)$ represents the semantic decoder with parameters χ and $C^{-1}_{\sigma}(\cdot)$ represents the channel decoder with σ parameters.

B. Architecture

Figure 1 presents the proposed architecture for the semantic communication system. It contains semantic coding and decoding modules, provided by pre-trained models. Also, the proposed systems have channel coding and decoding modules made of fully connected layers. The use of pre-trained models together with fully connected networks as channel encoders is one of the innovations of the proposed system.

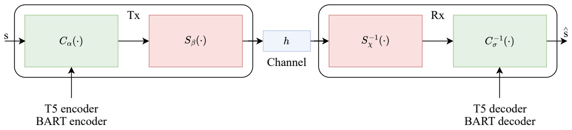

Fig. 1: Semantic communication system based on [13].

The semantic coding decoding steps are based on pre-trained Transformers models. Such models have separate encoder and decoder modules, which are trained together.

The channel coding and decoding steps were implemented using neural networks, as proposed in [9]. Figure 2 illustrates the channel encoding step using dense layers. Other channel encoding alternatives could be considered, as in [13].

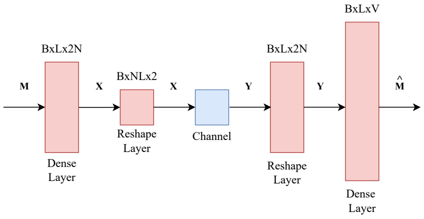

Fig. 2: Channel coding architecture based on [9].

C. Model Training

The system training proposal is summarized in Algorithm 1. The encoded signal $\mathbf{x}$ is obtained by the semantic coding of the symbols $\boldsymbol{s}$. Then, the semantic codes are transmitted over the communication channel. After that, channel and semantic decoding are applied to obtain the $\hat{\boldsymbol{s}}$ sentences.

To update the weights α, β, σ, χ, a good choice for the loss equation is cross entropy (CE), evaluated as

$$\mathcal{L}_{CE} = -\sum q(w_l)\log(p(w_l)) + (1 - q(w_l))\log(1 - p(w_l)), \tag{4}$$

where $q(w_l)$ is the probability that the l-th word, w_l, appears in the transmitted sentence $\mathbf{s}$, and $p(w_l)$ is the predicted probability that the i-th word, w_i, appears in the decoded sentence $\hat{\boldsymbol{s}}$.

The choice of the cross-entropy equation to calculate the error is due to its ability to measure the difference between two different probability distributions. The network, presented in Figure 1, is able to learn the distribution $q(w_l)$ given by the input sentence $\hat{\boldsymbol{s}}$, indicating that the syntax and the meaning of the word in the context are being learned by the network.

Algorithm 1 System-wide training algorithm.

```
Inicialização: Initial weights $\alpha, \beta, \sigma$ and $\chi$ from previous steps.
Inputs: Input tokens $\mathbf{s}$.
Transmition:
    $\mathbf{x} \leftarrow C_\alpha(S_\beta(\boldsymbol{s}))$,
    Transmits $\mathbf{x}$ over the communication channel.
Reception:
    Detects $\hat{\mathbf{y}}$ on the receiver,
    $\hat{\mathbf{x}} \leftarrow S^{-1}_\chi(C^{-1}_\sigma(\hat{\mathbf{y}}))$,
    Calculates $\mathcal{L}_{CE}$ by the Equation (4) comparing $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$,
    Trains $S_\alpha$, $C_\beta$, $C^{-1}_\sigma$ and $S^{-1}_\chi$ $\leftarrow$ Gradient Descent with $\mathcal{L}_{CE}$.
Outputs: Updated weights of networks $S_\alpha$, $C_\beta$, $C^{-1}_\sigma$ and $S^{-1}_\chi$.
```

An important emphasis must be given to training sequence-to-sequence models. First, such models create output sequences (or just sentences) from input sequences. In that way, the generation of output data occurs in a contextual manner; for example, in the context of this work, to generate a word, the model needs to know the previous word provided. This alternative differs from generalist auto-encoder models, which are trained by comparing entire sentences leaving the decoder. As BART [11] and T5 [12] models are sequence-to-sequence types, their training is done in a contextual manner.

III. Performance Metrics

For the study presented in this paper, involving transmission of text, the use of performance metrics such as BER (*Bit Error Rate*) is not justified. In transmitting semantic information, we must focus only on comparing the transmitted text with the received text. Thus, the bit error is irrelevant to the proposal's objective and does not accurately reflect the system's performance in the desired way. In fact, a literature survey on semantic communication systems shows that researchers are still looking for a better way to measure the performance of the proposed models. Two alternatives often used, as in [9] and [15], are BLEU (*Bilingual Evaluation Understudy*) and BERTScore, described next.

A. BLEU

The BLEU metric was first proposed in [16] to evaluate text translations. The value of this metric is given by the equivalence between text segments. Therefore, the metric is decomposed into n-grams for each group made up of one to four words. Such groupings are compared, and the metric is evaluated as follows:

$$\log(\text{BLEU}) = \min\left(1 - \frac{l_{\hat{s}}}{l_s}, 0\right) \sum_{n=1}^{N} u_n \log p_n, \tag{5}$$

where u_n are the weights of the n-grams and p_n is the n-gram score calculated according to

$$p_n = \frac{\sum_k \min(C_k(\hat{\boldsymbol{s}}), C_k(\boldsymbol{s}))}{\sum_k \min(C_k(\hat{\boldsymbol{s}})))}, \tag{6}$$

where $C_k(\cdot)$ is the frequency count for the k-th element in the n-gram.

The BLEU value varies from 0 to 1. For the proposed study, the metric is used to evaluate up to four grams. For the 1-gram (or unigram), the metric compares the frequency of each word in the decoded sentence with the original sentence. For the 2-gram (or bigram), this comparison uses groups of two words. For 3-gram, the comparison is between groups of three words, and for 4-gram, the comparison is between groups of four words.

In this way, if a group of words has a frequency of occurrence in the decoded sentence equivalent to the frequency in the original sentence, the BLEU value will be higher. It is important to note that this comparison considers the exact occurrence of each word or group of words in the text.

For better notation, the BLEU values for each n-gram will be written as BLEU-1, BLEU-2, BLEU-3, and BLEU-4 in accordance with the n-gram groupings.

B. BERTScore

As noted in the previous section, the BLEU metric only evaluates the equivalence between words within the same group of n-grams. In this way, the comparison between different words with the same meaning would give a bad result by the metric, but the sentences could have the same meaning. Thus, the evaluation of the transmission of semantics between different sentences would be compromised.

To overcome this difficult, some authors, such as Xie *et al.* in [9], started using the BERTScore metric. This metric was first proposed in [17] and considers the pre-trained contextual embeddings of the BERT model - a state-of-the-art natural language processing (NLP) model composed of several Transformers layers and trained in a bidirectional manner, as deeply explained in [18]. To calculate the metric, cosine similarity is considered, as in the expression

$$\text{BERTScore} = \frac{\boldsymbol{B}_{\boldsymbol{\Phi}}(\boldsymbol{s}) \cdot \boldsymbol{B}_{\boldsymbol{\Phi}}(\hat{\boldsymbol{s}})^T}{\left\|\boldsymbol{B}_{\boldsymbol{\Phi}}(\boldsymbol{s})\right\| \left\|\boldsymbol{B}_{\boldsymbol{\Phi}}(\hat{\boldsymbol{s}})\right\|}, \tag{7}$$

where $\boldsymbol{B}_{\boldsymbol{\Phi}}(\boldsymbol{s})$ represents the vectors extracted from the pre-trained BERT model, according to the input tokens $\boldsymbol{s}$. Note that this metric also varies from 0 to 1 to measure the similarity between the phrases $\boldsymbol{s}$ and $\hat{\boldsymbol{s}}$.

Since the BERT model was trained from millions of sentences, we can conclude that it learned the semantics coming from these input texts. Thus, the operator $\boldsymbol{B}_{\boldsymbol{\Phi}}(\cdot)$ must present close vectors if the semantic value between them is also close. This aspect differentiates the BERTScore metric from the BLEU, as the latter cannot identify semantic similarities between different sentences.

IV. Numerical Results

A traditional signal transmission scheme was considered as a baseline for comparison. This scheme comprises the Huffman algorithm for entropy coding, Turbo coding for channel coding, and 128-QAM modulation. For Turbo coding, polynomials of size 4 and a code rate equal to 0.5 were used. We also utilized 5 iterations and the maxlog decoding method for Turbo decoding. To implement this coding alternative, the NVIDIA Sionna library [19] was employed.

The choice of the T5 and BART models is justified because they both divide their architectures into encoding and decoding modules. As presented in references [11] and [12], BART and T5 models have different distilled versions with varying amounts of parameters.

To obtain the results, the smallest available versions of each of these models were considered. This alternative facilitates model training and demonstrates that even more compressed versions of pre-trained models can give good results for developing semantic-based communication systems. In that way, the T5-SMALL and BART-BASE versions were used. It is worth noting that T5-SMALL has 60.5M parameters, and BART-BASE has 139M parameters. Specifically for this work, the model created using T5-SMALL was called T5-SC, and the one created using BART-BASE was called BART-SC.

Table I summarizes the proposed semantic coding models, with the number of output units of each layer.

TABLE I: Semantic Communication summary architecture.

	Layer Name	Units	Activation Function
Transmitter (Encoder)	Transformer Encoder	128	Linear
	Dense	256	Relu
	Dense	16	Relu
Channel	AWGN	None	None
Receiver (Decoder)	Dense	16	ReLu
	Dense	256	ReLu
	Transformer Decoder	128	Softmax

For the proposed models T5-SC and BART-SC, it was necessary to jointly train the coding and decoding modules in just one epoch. It was observed that the loss function metrics during this training decreased as expected. This is justified by the pre-training of the T5 and BART models. For T5-SC, it was necessary to use a learning rate equal to 0.001, while for BART-SC, a learning rate equal to 0.0005 was considered. Such values were obtained experimentally by training the models using portions of the original training data.

For a better comparison of results, the proposed models T5-SC and BART-SC were trained under the same condition – AWGN channel at a signal-to-noise ratio (SNR) equal to 12dB.

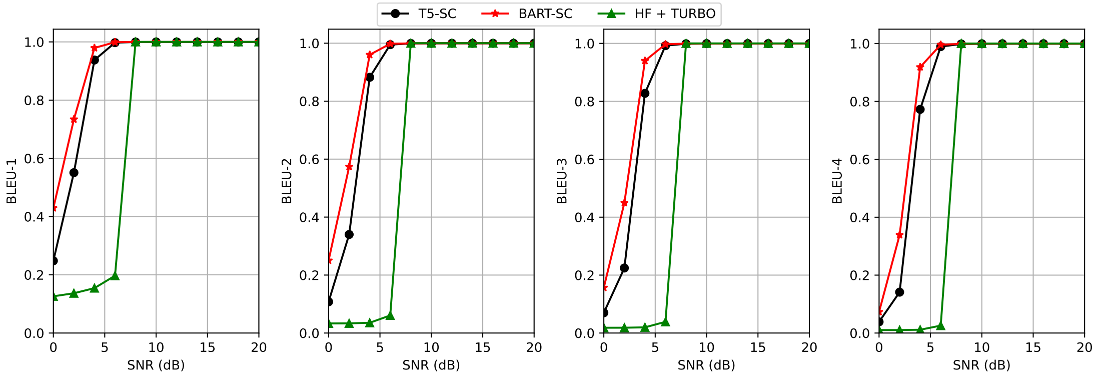

Fig. 3: BLEU results for AWGN only channel. The proposed models present better BLEU results in low SNR conditions.

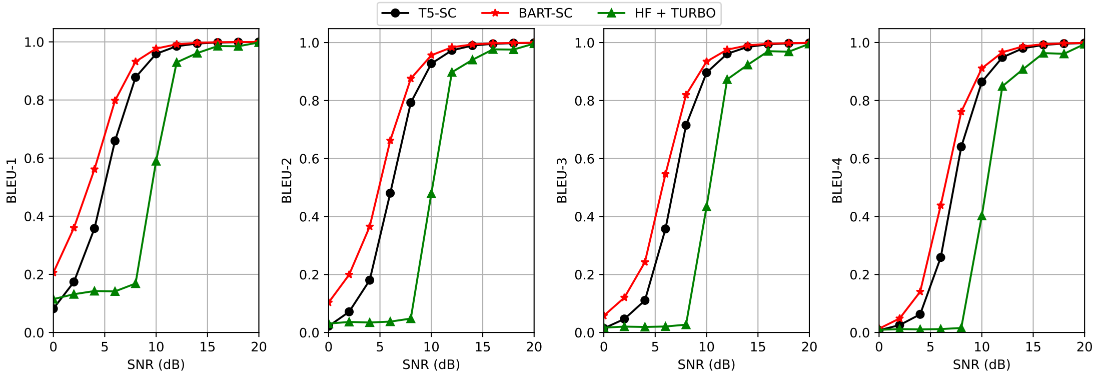

Fig. 4: BLEU results for Rayleigh channel. As expected, the Rayleigh fading causes a noticeable degradation to the results, but the proposed system maintains better results compared with the traditional alternative.

TABLE II: Sample sentences over Rayleigh fading channel for SNR = 12 dB.

	Sentence	BLEU-1	BERTScore
Transmitted sentence	**ii cooperation procedure (second reading)	-	-
Received, HF + TURBO	**ii cooperation procelure (dpb d rearttdt/sta u	0.50	0.81
Received, T5-SC	**ii cooperation procedure (second reading)	1.00	1.00
Received, BART-SC	**ii cooperation procedure (second reading)	1.00	1.00

The dataset used for training, validation, and testing is the proceedings of the European Parliament [20]. This dataset contains 2 million sentences and approximately 53 million words. To allow comparison with the results of [9], only sentences containing 4 to 30 words were used in our experiments.

Figure 3 presents the BLEU results in different n-grams (BLEU-1, BLEU-2, BLEU-3 and BLEU-4) under the AWGN channel conditions. We can see that for low SNR (poor channel conditions), the pre-trained models present better BLEU results than the traditional system. As expected, the BLEU metrics for larger n-grams present worse results for the same channel condition. This is due to the difficulty in observing the exact same group of words between the input sentence and the decoded sentence.

Figure 4 shows the same BLEU metrics, but now under the Rayleigh channel, having been trained under the AWGN channel conditions (as discussed above). As expected, at the same SNR level, the results under Rayleigh fading are worse compared with the AWGN-only channel. Even so, the results for the proposed models are still better than the traditional coding case. This is especially seen at low SNR levels.

For larger blocks of words, such as 3-gram and 4-gram, the decoding of the T5-SC and BART-SC semantic models was better for low SNR levels, such as 10dB, as can be seen in the Figures 3 and 4. In this way, the decoding promoted in a contextual way in sequence-to-sequence models presents results of words that are related to each other, resulting in good decoding of textual blocks.

Low BLEU values for some decoding cases may not have been caused by word choices in similar but different contexts to the original sentence. This fact indicates a syntactic error highlighted by the BLEU metric but may not mean a semantic error.

Figure 5 presents the BERTScore results for the proposed models considering a Rayleigh channel. For the results in Figure 5, the improved performance of the T5-SC and BART-SC models is noted when compared to the traditional coding models. Using BERTScore metric, we can assume that at low

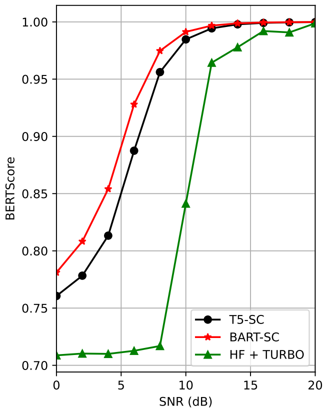

Fig. 5: BERTScore results for Rayleigh channel.

SNR levels, such as 6dB, the decoded text could already be understandable by the receiver, with a BERTScore above 0.80.

Moreover, with the aim of comparing the results, Table II presents the same sentence in the 12dB condition on a channel with the presence of Rayleigh fading for the four models obtained.

V. Conclusions

In this paper, we propose two semantic communication models using pre-trained language models. Due to the architecture of each model, T5 and BART models were chosen. As a comparison, a traditional encoding model based on Huffman and Turbo coding was used.

Numerical comparisons demonstrate the advantage of using pre-trained language models T5 and BART. For the different blocks of words in the BLEU metric, the semantic models presented a better performance, especially under low SNR. When considering the BERTScore metric, there are also better results for semantic cases, indicating that the decoded sentences present a better response for human judgment. It is also important to note that the semantic models were trained just in one condition (AWGN channel and SNR equals 12 dB) but demonstrated good results in the presence of Rayleigh fading. Also, the proposed models used just one epoch of training. This point indicates a good perspective of using those models to encode texts with different base knowledge.

Furthermore, this proposal indicates a new perspective on the objective of communication - it would not be necessary to obtain the message decoded exactly, but only something close and with the same semantic notion. Therefore, new metrics such as BERTScore must be considered.

The good results obtained in this research open up the option to use pre-trained language models on other data sources or with other communication system architectures.

Aknowledgements

This work was supported by the National Center for High Performance Processing in São Paulo (CENAPAD-SP) with the availability of GPUs for training and testing the proposed models. Also, we are grateful for the financial support from CAPES and CNPq.

References

[1] Shannon, Claude Elwood, *A mathematical theory of communication*, The Bell system technical journal, v. 27, n. 3, pp. 379–423, Nokia Bell Labs, 1948.

[2] Shi, Guangming and Gao, Dahua and Song, Xiaodan and Chai, Jingxuan and Yang, Minxi and Xie, Xuemei and Li, Leida and Li, Xuyang, *A new communication paradigm: from bit accuracy to semantic fidelity*, arXiv preprint arXiv:2101.12649, 2021.

[3] C.Shannon,S.E,W.Weaver,, R. Blahut, and B. Hajek, *The Mathematical Theory of Communication,* ser. Illini books. University of Illinois Press, 1963.

[4] Carnap, Rudolf and Bar-Hillel, Yehoshua and others *An outline of a theory of semantic information*, Research Laboratory of Electronics, Massachusetts Institute of Technology, 1952.

[5] Wheeler, Dylan and Natarajan, Balasubramaniam, *Engineering semantic communication: A survey*, IEEE Access, v. 11, pp. 13965–13995, IEEE, 2023.

[6] Delgado-Frias, Jose G and Moore, Will R, *A semantic network architecture for artificial intelligence processing*, IEEE International Workshop on Tools for Artificial Intelligence, pp. 162–163, IEEE Computer Society, 1989.

[7] Zhou, Qingyang and Li, Rongpeng and Zhao, Zhifeng and Peng, Chenghui and Zhang, Honggang, *Semantic communication with adaptive universal transformer*, IEEE Wireless Communications Letters, v. 11, n. 3, pp. 453–457, IEEE, 2021.

[8] Letaief, Khaled B and Chen, Wei and Shi, Yuanming and Zhang, Jun and Zhang, Ying-Jun Angela *The roadmap to 6G: AI empowered wireless networks*, IEEE communications magazine, v. 57, n. 8, pp. 84–90, IEEE, 2019.

[9] Xie, Huiqiang and Qin, Zhijin and Li, Geoffrey Ye and Juang, Biing-Hwang, *Deep learning enabled semantic communication systems*, *IEEE Transactions on Signal Processing*, v. 69, pp. 2663–2675, IEEE, 2021.

[10] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia, *Attention is all you need*, Advances in neural information processing systems, v. 30, 2017.

[11] Mike Lewis and Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, Luke Zettlemoyer, *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, CoRR, v. abs/1910.13461, http://arxiv.org/abs/1910.13461, arxiv, Oct 2019.

[12] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, Journal of Machine Learning Research, v. 21, n. 140, pp 1–67, http://jmlr.org/papers/v21/20-074.html, 2020.

[13] Lee, Ju-Hyung and Lee, Dong-Ho and Sheen, Eunsoo and Choi, Thomas and Pujara, Jay *Seq2Seq-SC: End-to-end semantic communication systems with pre-trained language model*, 2023 57th Asilomar Conference on Signals, Systems, and Computers, pp. 260–264, IEEE, 2023.

[14] Xie, Huiqiang and Qin, Zhijin, *A lite distributed semantic communication system for internet of things*, IEEE Journal on Selected Areas in Communications, v. 39, pp 142–153, IEEE, 2020.

[15] Peng, Xiang and Qin, Zhijin and Huang, Danlan and Tao, Xiaoming and Lu, Jianhua and Liu, Guangyi and Pan, Chengkang *A robust deep learning enabled semantic communication system for text*, GLOBECOM 2022-2022 IEEE Global Communications Conference, pp. 2704–2709, IEEE, 2022.

[16] Kishore Papineni and Salim Roukos and Todd Ward and Wei-jing Zhu *BLEU: a Method for Automatic Evaluation of Machine Translation*, pp. 311–318, 2002.

[17] Tianyi Zhang* and Varsha Kishore* and Felix Wu* and Kilian Q. Weinberger and Yoav Artzi *BERTScore: Evaluating Text Generation with BERT*, International Conference on Learning Representations, https://openreview.net/forum?id=SkeHuCVFDr, 2020.

[18] Jacob Devlin, Ming-Wei Chang and, Kenton Lee and Kristina Toutanova, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, CoRR, v. abs/1810.04805, http://arxiv.org/abs/1810.04805, 2018.

[19] Hoydis, Jakob and Cammerer, Sebastian and Aoudia, Fayçal Ait and Vem, Avinash and Binder, Nikolaus and Marcus, Guillermo and Keller, Alexander, *Sionna: An open-source library for next-generation physical layer research*, arXiv preprint arXiv:2203.11854, 2022.

[20] Koehn, Philipp *Europarl: A parallel corpus for statistical machine translation*, Proceedings of machine translation summit x: papers, pp. 79–86, 2005.