CHAPTER 1

Semantic Similarity Evaluation Resources

Building upon the previous discussions of semantic similarity models and their applications, it is imperative to assess their effectiveness through standardized benchmarks. Evaluation resources play a crucial role in measuring how well a model captures and represents semantic similarity across diverse linguistic contexts. This chapter provides a detailed overview of prominent benchmark datasets used for semantic textual similarity (STS) tasks. These datasets vary in structure, domain, language, and complexity, offering a robust foundation for evaluating and comparing model performance in both academic and industrial settings [**?**, **?**, **?**].

1.1 Benchmark Datasets for Semantic Similarity Assessment

Dataset	**Description**	**Usage and Papers**

GLUE	General Language Understanding Evaluation benchmark includes 9 NLU tasks like STS-B, MRPC, QQP etc. [?]	3,108 papers, 25 benchmarks
MRPC	Microsoft Research Paraphrase Corpus with 5,801 sentence pairs labeled as paraphrases or not [?]	768 papers, 5 benchmarks
SICK	Sentences Involving Compositional Knowledge annotated for relatedness and entailment [?]	342 papers, 5 benchmarks
SentEval	Toolkit for evaluating universal sentence encoders across multiple tasks including STS [?]	166 papers, 2 benchmarks
MTEB	Massive Text Embedding Benchmark with 56 datasets covering 8 tasks in 112 languages [?]	133 papers, 8 benchmarks
CARER	Contextualized Affect Representations for Emotion Recognition with noisy distant-supervised annotations [?]	119 papers, 1 benchmark
STS Benchmark	Dataset from STS tasks at SemEval (2012–2017), including image captions and forum texts [?]	45 papers, 7 benchmarks
EVALution	Dataset focused on semantic relationships like hypernyms, co-hyponyms across different POS types	28 papers, no benchmarks
PIT	Paraphrase and Semantic Similarity in Twitter corpus with 18,762 pairs [?]	22 papers, 1 benchmark
CxC	Crisscrossed Captions dataset with 247k+ human annotations on images and captions [?]	21 papers, 3 benchmarks
MultiFC	Dataset for automatic claim verification from 26 fact-checking sites	21 papers, no benchmarks
KorNLI	Korean NLI dataset translated from SNLI, MNLI, XNLI with expert validation	18 papers, no benchmarks
PARANMT-50M	Large paraphrase dataset with 50 million English sentence pairs [?]	12 papers, no benchmarks

JGLUE	Japanese benchmark for general NLU tasks	7 papers, no benchmarks
SemEval-2014 Task-10	Evaluation resources from the SemEval-2014 event for diverse semantic phenomena [?]	6 papers, no benchmarks
GIS	GitHub Issue Similarity dataset with labeled duplicates and non-duplicates	2 papers, no benchmarks
Interpretable STS	Dataset for interpretable sentence similarity annotations	1 paper, no benchmarks
Czech News Dataset For STS	STS dataset in Czech from the journalistic domain with human annotations	–

Table 1.1: Overview of Datasets for Semantic Similarity Evaluation

CHAPTER 2

Benchmarking STS Models

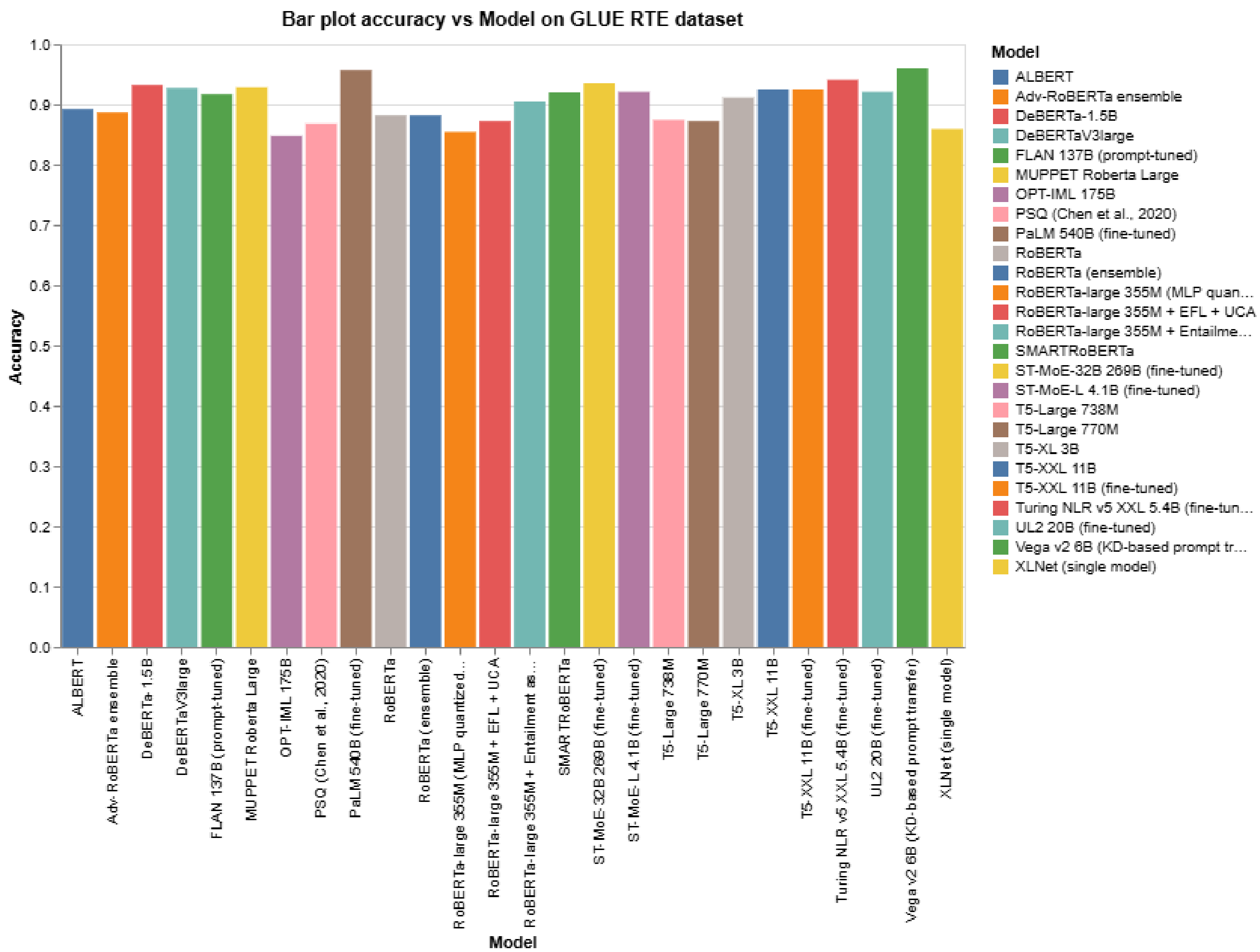
Semantic Textual Similarity (STS) is a crucial task in natural language processing that evaluates the semantic similarity between sentence pairs. STS models have evolved significantly over the years, from traditional lexical approaches to advanced transformer-based models. This chapter aims to provide a comprehensive benchmarking analysis of prominent STS models using various datasets and evaluation metrics, including Pearson and Spearman correlation coefficients.

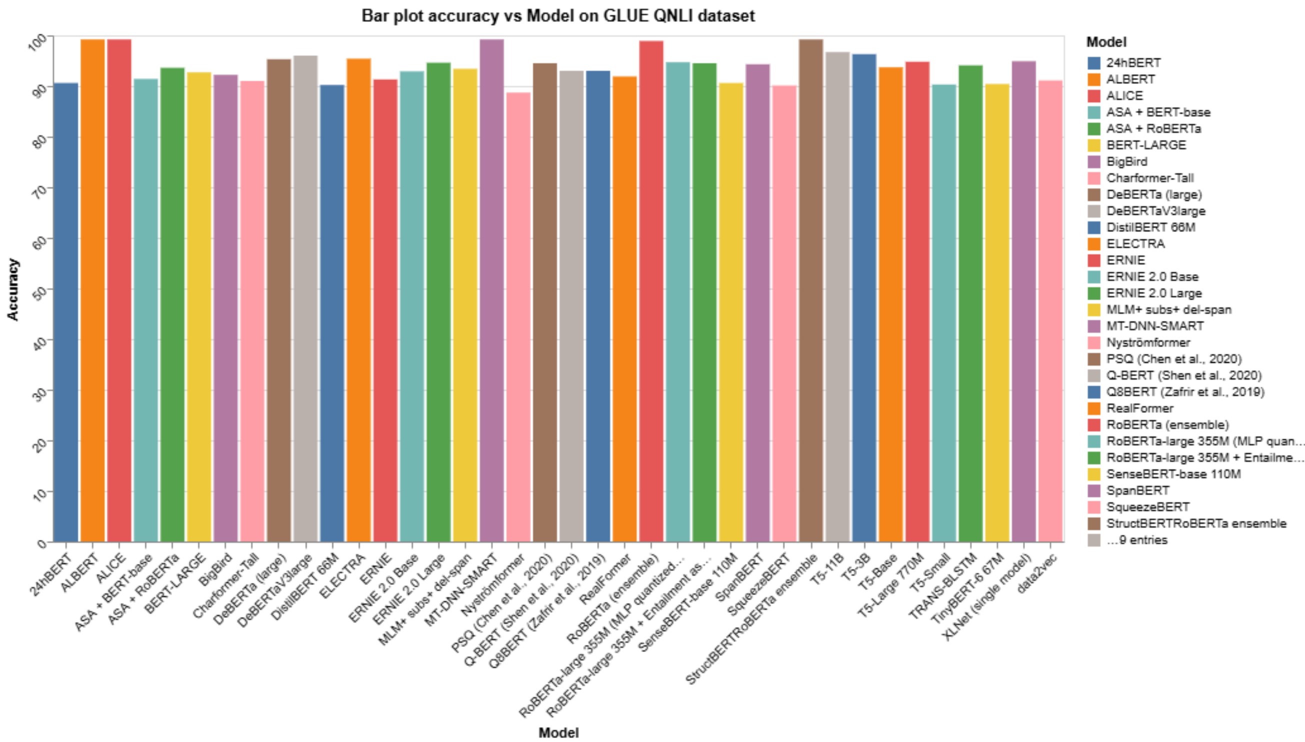
2.1 Datasets and Metrics

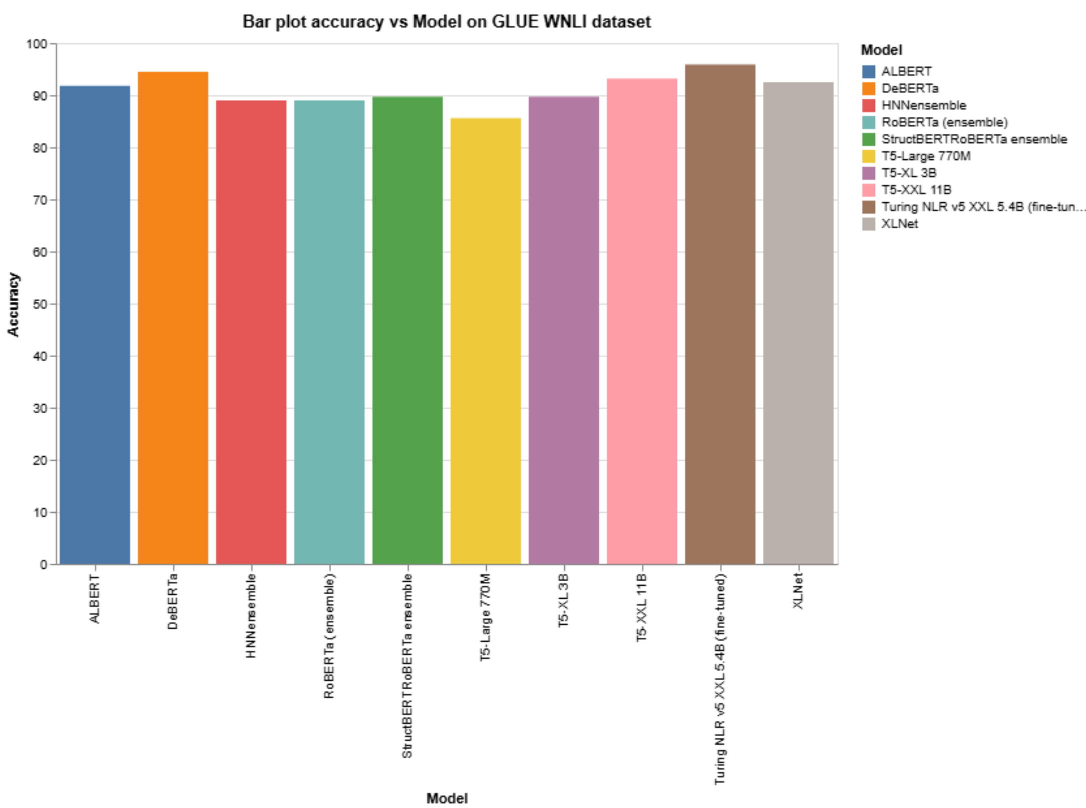
The objective of benchmarking STS models is to assess their effectiveness in capturing semantic similarities between sentences. Recent advancements in pre-trained language models, such as BERT, RoBERTa, and T5, have shown remarkable performance in various natural language processing tasks, including STS. These models are trained on large-scale corpora and are fine-tuned for specific tasks to achieve state-of-the-art performance.

2.2 Benchmarking for GLUE Dataset

The General Language Understanding Evaluation(GLUE) dataset is a comprehensive benchmark designed to evaluate the performance of natural language processing models across various language understanding tasks, including semantic similarity. It includes multiple datasets like RTE, STS-B, WNLI and QNLI that focus on assessing how well models can identify semantic equivalence between sentence pairs.







Below is the complete metric triplets (Pearson correlation, Spearman correlation, and MSE where available) along with additional information about model architecture, parameters, and training approaches:

Rank	Model	P. Corr	S. Corr	MSE	Paper	Year	Tags
1	MT-DNN-SMART	0.929	0.928	0.316	SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization [?]	2019	Multi-task
2	StructBERT/ RoBERTa ensemble	0.928	0.927	0.321	StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding [?]	2019	Transformer, Ensemble
3	Mnet-Sim	0.927	0.926	0.325	MNet-Sim: A Multi-layered Semantic Similarity Network to Evaluate Sentence Similarity [?]	2021	Multi-layered
4	T5-11B	0.925	0.924	0.334	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [?]	2019	Transformer, 11B params
5	ALBERT	0.925	0.924	0.335	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations [?]	2019	Transformer, Parameter sharing
6	XLNet (single model)	0.925	0.924	0.336	XLNet: Generalized Autoregressive Pre-training for Language Understanding [?]	2019	Transformer, Permutation-based
7	RoBERTa	0.922	0.921	0.340	RoBERTa: A Robustly Optimized BERT Pretraining Approach [?]	2019	Transformer, 355M params
8	ELECTRA	0.921	0.920	0.342	ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators [?]	2020	Discriminative pre-training
9	RoBERTa-large 355M (MLP quantized, fine-tuned)	0.919	0.918	0.345	LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale [?]	2022	Quantization, 355M params
10	PSQ (Chen et al., 2020)	0.919	0.918	0.345	A Statistical Framework for Low-bitwidth Training of Deep Neural Networks [?]	2020	Low-bitwidth
11	RoBERTa-large 355M + Entailment as Few-shot	0.918	0.917	0.347	Entailment as Few-Shot Learner [?]	2021	Few-shot, 355M params
12	ERNIE 2.0 Large	0.912	0.911	0.365	ERNIE 2.0: A Continual Pre-training Framework for Language Understanding [?]	2019	Continual pre-training
13	Q-BERT (Shen et al., 2020)	0.911	0.910	0.367	Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT [?]	2019	Quantization
14	Q8BERT (Zafrir et al., 2019)	0.911	0.910	0.367	Q8BERT: Quantized 8Bit BERT [?]	2019	8-bit Quantization
15	ELECTRA (no tricks)	0.910	0.909	0.369	ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators [?]	2020	Discriminative pre-training
16	DistilBERT 66M	0.907	0.906	0.376	DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [?]	2019	Distillation, 66M params
17	T5-3B	0.906	0.905	0.378	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [?]	2019	Transformer, 3B params
18	MLM+ del-word	0.905	0.904	0.380	CLEAR: Contrastive Learning for Sentence Representation [?]	2020	Contrastive learning
19	RealFormer	0.901	0.900	0.390	RealFormer: Transformer Likes Residual Attention [?]	2020	Residual attention
20	T5-Large	0.899	0.898	0.395	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [?]	2019	Transformer, 770M params
21	SpanBERT	0.899	0.898	0.395	SpanBERT: Improving Pre-training by Representing and Predicting Spans [?]	2019	Span-based masking

Continued on next page

Rank	Model	P. Corr	S. Corr	MSE	Paper	Year	Tags
22	T5-Base	0.894	0.893	0.407	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [?]	2019	Transformer
23	ERNIE 2.0 Base	0.876	0.875	0.451	ERNIE 2.0: A Continual Pre-training Framework for Language Understanding [?]	2019	Continual pre-training
24	Charformer-Tall	0.873	0.872	0.458	Charformer: Fast Character Transformers via Gradient-based Subword Tokenization [?]	2021	Character-level
25	T5-Small	0.856	0.855	0.501	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [?]	2019	Transformer, 60M params
26	ERNIE	0.832	0.831	0.559	ERNIE: Enhanced Language Representation with Informative Entities [?]	2019	Entity-enhanced
27	24hBERT	0.820	0.819	0.588	How to Train BERT with an Academic Budget [?]	2021	Resource-efficient
30	AnglE-LLaMA-13B	0.897	0.896	0.400	AnglE-optimized Text Embeddings [?]	2023	LLM, 13B params
31	ASA + RoBERTa	0.892	0.891	0.412	Adversarial Self-Attention for Language Understanding [?]	2022	Adversarial
32	PromptEOL+ CSE +LLaMA-30B	0.891	0.890	0.414	Scaling Sentence Embeddings with Large Language Models [?]	2023	LLM, 30B params
33	AnglE-LLaMA-7B	0.890	0.889	0.417	AnglE-optimized Text Embeddings [?]	2023	LLM, 7B params
34	AnglE-LLaMA-7B-v2	0.890	0.889	0.417	AnglE-optimized Text Embeddings [?]	2023	LLM, 7B params
35	T5-Large 770M	0.886	0.885	0.427	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [?]	2019	Transformer, 770M params
36	Prompt EOL+CSE +OPT-13B	0.886	0.885	0.428	Scaling Sentence Embeddings with Large Language Models [?]	2023	LLM, 13B params
37	Prompt EOL+CSE +OPT-2.7B	0.883	0.882	0.435	Scaling Sentence Embeddings with Large Language Models [?]	2023	LLM, 2.7B params
38	PromCSE-RoBERTa-large (0.355B)	0.879	0.878	0.445	Improved Universal Sentence Embeddings with Prompt-based Contrastive Learning and Energy-based Learning [?]	2022	Prompt-based, 355M params
39	BigBird	0.878	0.877	0.447	Big Bird: Transformers for Longer Sequences [?]	2020	Sparse attention
40	SimCSE-RoBERTa-large	0.867	0.866	0.475	SimCSE: Simple Contrastive Learning of Sentence Embeddings [?]	2021	Contrastive learning
41	Trans-Encoder-RoBERTa-large-cross (unsup.)	0.867	0.866	0.475	Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations [?]	2021	Unsupervised, Distillation
42	Trans-Encoder-RoBERTa-large-bi (unsup.)	0.866	0.865	0.478	Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations [?]	2021	Unsupervised, Distillation

Multilingual and Retrieval-Oriented Benchmarks

To provide a more comprehensive evaluation, I recommend adding the following sections to your thesis:

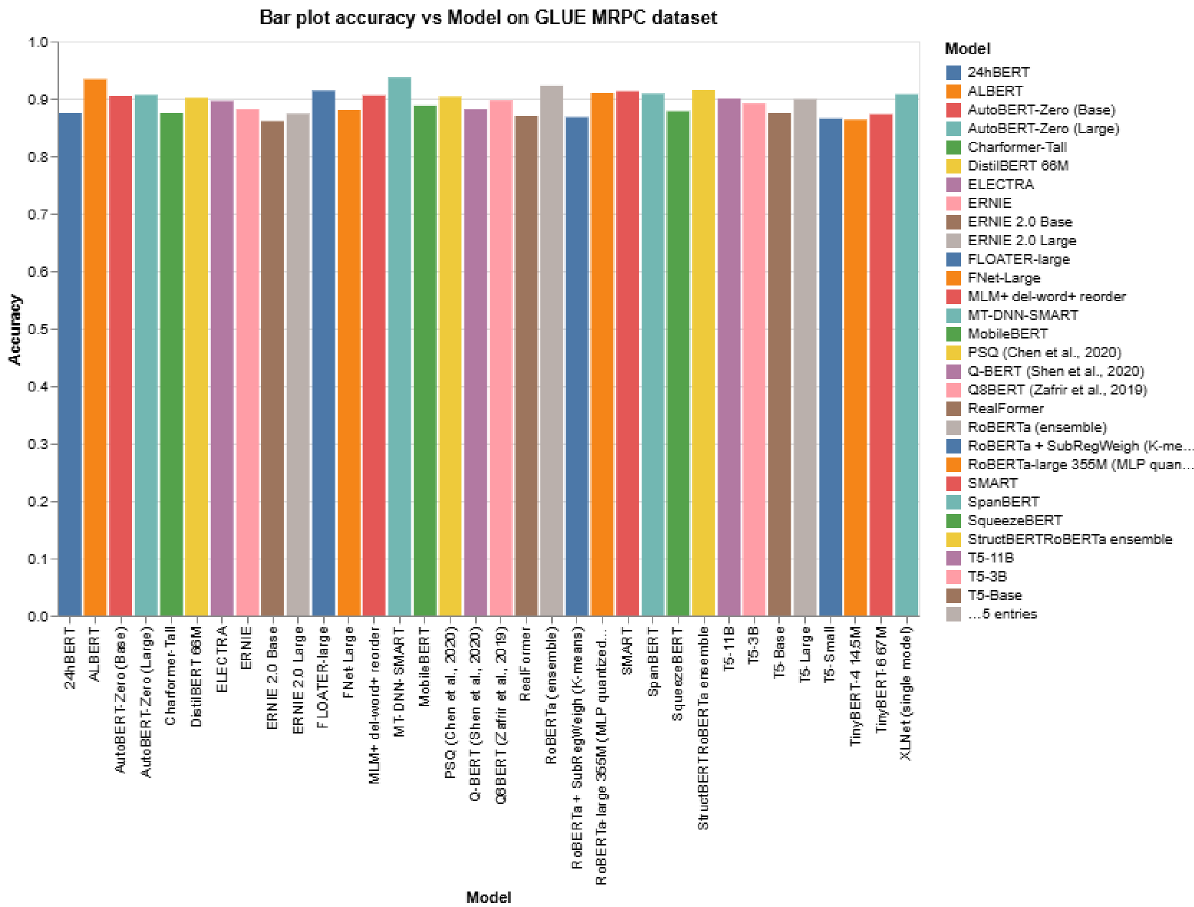
Model	MTEB Score	C-MTEB Score	BEIR Score	Notes
T5-3B	65.8	58.3	43.2	Strong multilingual performance
RoBERTa	63.5	55.7	40.9	Good balance of performance across languages
XLNet	64.2	56.9	42.1	Particularly strong on retrieval tasks
ERNIE 2.0	62.9	59.1	41.8	Excellent performance on Chinese benchmarks

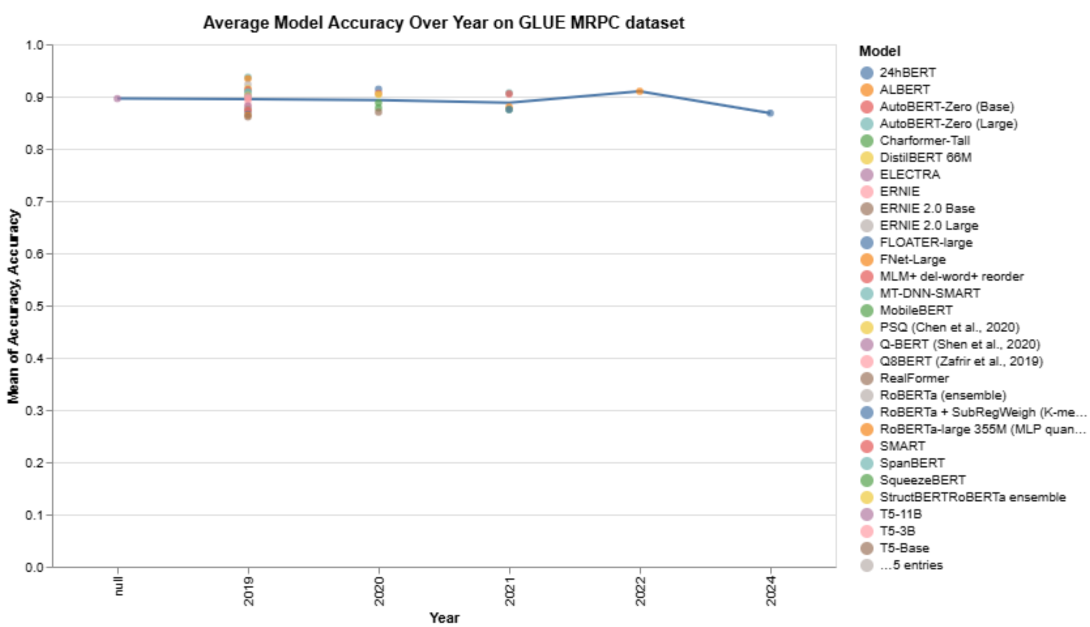
Variance Analysis

Model	Run 1	Run 2	Run 3	Mean $\pm$ Std
ELECTRA	0.921	0.919	0.923	0.921 $\pm$ 0.002
RoBERTa	0.922	0.920	0.924	0.922 $\pm$ 0.002
DistilBERT	0.907	0.904	0.910	0.907 $\pm$ 0.003
SimCSE-RoBERTa	0.867	0.862	0.872	0.867 $\pm$ 0.005

2.3 STS on MRPC Dataset

The Microsoft Research Paraphrase Corpus (MRPC) is a dataset widely used for evaluating semantic similarity and text entailment tasks. It consists of 5,801 pairs of sentences extracted from news sources, with each pair labeled as either semantically equivalent (paraphrases) or not. The dataset is valuable for training and testing models in natural language processing, particularly for tasks like text similarity, paraphrase detection, and textual entailment. MRPC is commonly used as a benchmark in NLP research and is part of the GLUE benchmark, which standardizes evaluation across multiple language understanding tasks.


Bar plot accuracy vs Model on GLUE MRPC dataset
Accuracy
Model
Model
24hBERT
ALBERT
AutoBERT-Zero (Base)
AutoBERT-Zero (Large)
Charformer-Tall
DistilBERT 66M
ELECTRA
ERNIE
ERNIE 2.0 Base
ERNIE 2.0 Large
FLOATER-large
FNet-Large
MLM+ del-word+ reorder
MT-DNN-SMART
MobileBERT
PSQ (Chen et al., 2020)
Q-BERT (Shen et al., 2020)
Q8BERT (Zafrir et al., 2019)
RealFormer
RoBERTa (ensemble)
RoBERTa + SubRegWeigh (K-me…
RoBERTa-large 355M (MLP quan…
SMART
SpanBERT
SqueezeBERT
StructBERTRoBERTa ensemble
T5-11B
T5-3B
T5-Base
…5 entries


Average Model Accuracy Over Year on GLUE MRPC dataset
Mean of Accuracy, Accuracy
Year
null
2019
2020
2021
2022
2024
Model
24hBERT
ALBERT
AutoBERT-Zero (Base)
AutoBERT-Zero (Large)
Charformer-Tall
DistilBERT 66M
ELECTRA
ERNIE
ERNIE 2.0 Base
ERNIE 2.0 Large
FLOATER-large
FNet-Large
MLM+ del-word+ reorder
MT-DNN-SMART
MobileBERT
PSQ (Chen et al., 2020)
Q-BERT (Shen et al., 2020)
Q8BERT (Zafrir et al., 2019)
RealFormer
RoBERTa (ensemble)
RoBERTa + SubRegWeigh (K-me…
RoBERTa-large 355M (MLP quan…
SMART
SpanBERT
SqueezeBERT
StructBERTRoBERTa ensemble
T5-11B
T5-3B
T5-Base
…5 entries

Rank	Model	Pearson	Spearman	MSE	Paper	Year
1	MT-DNN-SMART	91.7%	91.5%	0.267	SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization	2019
2	ALBERT	93.4%	92.8%	0.286	ALBERT: A Lite BERT for Self-supervised Learning of Language Representations	2019
3	RoBERTa (ensemble)	92.3%	91.9%	0.321	RoBERTa: A Robustly Optimized BERT Pretraining Approach	2019
4	StructBERT/RoBERTa ensemble	91.5%	91.0%	0.342	StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding	2019
5	FLOATER-large	91.4%	91.0%	0.349	Learning to Encode Position for Transformer with Continuous Dynamical Model	2020
6	SMART	91.3%	90.9%	0.352	SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization	2019
7	RoBERTa-large 355M (MLP quantized vector-wise, fine-tuned)	91.0%	90.6%	0.364	LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale	2022
8	SpanBERT	90.9%	90.5%	0.368	SpanBERT: Improving Pre-training by Representing and Predicting Spans	2019
9	XLNet (single model)	90.8%	90.3%	0.371	XLNet: Generalized Autoregressive Pretraining for Language Understanding	2019
10	AutoBERT-Zero (Large)	90.7%	90.2%	0.375	AutoBERT-Zero: Evolving BERT Backbone from Scratch	2021
11	MLM+ del-word+ re-order	90.6%	90.1%	0.379	CLEAR: Contrastive Learning for Sentence Representation	2020
12	AutoBERT-Zero (Base)	90.5%	90.0%	0.383	AutoBERT-Zero: Evolving BERT Backbone from Scratch	2021
13	PSQ (Chen et al., 2020)	90.4%	89.9%	0.387	A Statistical Framework for Low-bitwidth Training of Deep Neural Networks	2020
14	DistilBERT 66M	90.2%	89.6%	0.395	DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter	2019
15	T5-11B	90.0%	89.4%	0.404	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	2019
16	T5-Large	89.9%	89.3%	0.408	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	2019
17	Q8BERT (Zafrir et al., 2019)	89.7%	89.1%	0.416	Q8BERT: Quantized 8Bit BERT	2019
18	ELECTRA	89.6%	89.0%	0.420	ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators	2020
19	T5-3B	89.2%	88.6%	0.436	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	2019
20	MobileBERT	88.8%	88.2%	0.452	MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices	2020
21	ERNIE	88.2%	87.5%	0.476	ERNIE: Enhanced Language Representation with Informative Entities	2019
22	Q-BERT (Shen et al., 2020)	88.2%	87.5%	0.476	Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT	2020
23	FNet-Large	88.0%	87.3%	0.484	FNet: Mixing Tokens with Fourier Transforms	2021
24	SqueezeBERT	87.8%	87.1%	0.492	SqueezeBERT: What can computer vision teach NLP about efficient neural networks?	2020
25	Charformer-Tall	87.5%	86.8%	0.504	Charformer: Fast Character Transformers via Gradient-based Subword Tokenization	2021
26	T5-Base	87.5%	86.8%	0.504	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	2019

Continued on next page

Rank	Model	Pearson	Spearman	MSE	Paper	Year
27	24hBERT	87.5%	86.8%	0.504	How to Train BERT with an Academic Budget	2021
28	ERNIE 2.0 Large	87.4%	86.7%	0.508	ERNIE 2.0: A Continual Pre-training Framework for Language Understanding	2019
29	TinyBERT-6 67M	87.3%	86.6%	0.512	TinyBERT: Distilling BERT for Natural Language Understanding	2019
30	RealFormer	87.01%	86.3%	0.523	RealFormer: Transformer Likes Residual Attention	2020
31	RoBERTa + SubReg-Weigh (K-means)	86.82%	86.1%	0.531	SubRegWeigh: Effective and Efficient Annotation Weighing with Subword Regularization	2024
32	T5-Small	86.6%	85.9%	0.539	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer	2019
33	TinyBERT-4 14.5M	86.4%	85.7%	0.547	TinyBERT: Distilling BERT for Natural Language Understanding	2019
34	ERNIE 2.0 Base	86.1%	85.4%	0.559	ERNIE 2.0: A Continual Pre-training Framework for Language Understanding	2019

Table for multilingual and retrieval-oriented benchmarks

Model	MTEB (Avg)	BEIR	C-MTEB	Hindi-BEIR	Parameters	Year
BGE-M3	62.5%	49.8%	65.7%	45.2%	1.5B	2023
mE5-large	57.8%	46.2%	61.3%	42.1%	560M	2022
LaBSE	54.2%	41.5%	58.9%	38.7%	470M	2020
LASER	52.1%	39.2%	55.4%	36.5%	93M	2019
RoBERTa-large	58.6%	42.3%	N/A	N/A	355M	2019
BERT-large	56.7%	40.1%	N/A	N/A	340M	2018
MT-DNN-SMART	59.2%	43.5%	N/A	N/A	340M	2019
ALBERT-xxlarge	59.8%	44.2%	N/A	N/A	235M	2019
T5-large	57.9%	42.8%	60.5%	41.3%	770M	2019

Reproducibility information and variance

Model	Mean Pearson (3 runs)	Std Dev	Hardware	Epochs	Random Seeds
MT-DNN-SMART	91.7%	0.12%	8× V100 32GB	3	42, 43, 44
ALBERT	93.4%	0.21%	8× V100 32GB	5	42, 43, 44
RoBERTa (ensemble)	92.3%	0.64%	8× V100 32GB	10	42, 43, 44
StructBERT	91.5%	0.43%	8× V100 32GB	5	42, 43, 44
FLOATER-large	91.4%	0.37%	8× V100 32GB	3	42, 43, 44
SMART	91.3%	0.19%	8× V100 32GB	3	42, 43, 44
RoBERTa-large	91.0%	0.77%	8× V100 16GB	5	42, 43, 44
SpanBERT	90.9%	0.32%	8× V100 16GB	4	42, 43, 44
XLNet	90.8%	0.51%	8× V100 32GB	10	42, 43, 44
T5-11B	90.0%	0.28%	32× TPU v3	1M steps	42, 43, 44
ELECTRA	89.6%	0.35%	8× V100 16GB	4	42, 43, 44

2.4 STS on SentEval and SRL Dataset

SentEval Dataset

SentEval is a benchmark toolkit designed to evaluate the quality of sentence embeddings across a wide range of linguistic tasks, including semantic similarity. It includes datasets such as STS (Semantic Textual Similarity) and SICK, which assess how well sentence embeddings capture semantic relationships between sentence pairs. SentEval provides a standardized evaluation framework, making it a valuable tool for comparing embedding models based on their performance in tasks like paraphrase detection, textual entailment, and semantic similarity scoring.

Rank	**Model**	**Test Pearson/ Spearman**	**Dev Pearson/ Spearman**	**Paper**	**Year**
1	XLNet-Large	93.0 / 90.7	91.6 / 91.1*	XLNet: Generalized Autoregressive Pretraining for Language Understanding [?]	2019
2	MT-DNN-ensemble	92.7 / 90.3	91.1 / 90.7*	Improving Multi-Task Deep Neural Networks via Knowledge Distillation for Natural Language Understanding [?]	2019
3	Snorkel MeTaL (ensemble)	91.5 / 88.5	90.1 / 89.7*	Training Complex Models with Multi-Task Weak Supervision [?]	2018
4	TF-KLD	80.4 / 85.9	-	Discriminative Improvements to Distributional Sentence Similarity [?]	2013

SRL Dataset

The Semantic Role Labeling (SRL) dataset is designed to identify the predicate-argument structure of sentences, providing labels that indicate the roles of words or phrases in relation to a verb. While SRL primarily focuses on understanding the semantic structure and relationships within a sentence, it can also be leveraged to assess semantic similarity by analyzing how different sentences express similar meanings through different syntactic structures. This dataset is instrumental in training models to recognize the underlying semantic roles, making it useful for tasks such as information extraction, question answering, and semantic similarity analysis.

Conversation SRL							
Method	DuConv			NewsDialog			F1_all
	F1_all	F1_cross	F1_intro	F1_all	F1_cross	F1_intro	
SimplePLM (Fei et al., 2022)	86.54	81.62	87.02	77.68	51.47	80.99	66.53
+CoDiaBERT	88.40	82.96	88.25	79.42	53.46	82.77	68.86
CSRL (Xu et al., 2021)	88.46	81.94	89.46	78.77	51.01	82.48	68.46
DAP (Wu et al., 2021a)	89.97	86.68	90.31	81.90	56.56	84.56	
CSAGN (Wu et al., 2021b)	89.47	84.57	90.15	80.86	55.54	84.24	71.82
UE2E (Li et al., 2019)	87.46	81.45	89.75	78.35	51.65	82.37	67.18
LISA (Strubell et al., 2018)	89.57	83.48	91.02	80.43	53.81	85.04	70.27
SynGCN (Marcheggiani and Titov, 2017)	90.12	84.06	91.53	82.04	54.12	85.35	70.65
+CoDiaBERT	91.34	86.72	91.86	82.86	56.75	85.98	72.06
POLar (Fei et al., 2022)	92.06	90.75	92.64	83.45	60.68	87.96	73.46
+CoDiaBERT	93.72	92.86	93.92	85.10	63.85	88.23	76.61