# Contents

# Abstract

The domain of semantic textual similarity (STS) has witnessed significant advancements since 2021, driven by developments in transformer architectures, contrastive learning, and domain-specific STS solutions. This survey presents a comprehensive and current overview of the progress in semantic similarity, systematically categorizing the advancements into six key areas: (1) Transformer-based models, (2) Contrastive learning approaches, (3) Domain-specific models, (4) Multi-modal models, (5) Graph-based models, and (6) Knowledge-enhanced models. Transformer architectures, including FarSSiBERT and DeBERTa-v3, along with contrastive learning models such as AspectCSE, have established new benchmarks in STS tasks in recent years. Domain-specific models, such as CXR-BERT for medical text and Financial-STS for finance, underscore the adaptability of STS techniques in specialized domains through the integration of domain-specific pretrained representations. Multi-modal STS models that incorporate both text and auxiliary modalities such as visual or audio data provide additional dimensions for capturing semantic similarity. Furthermore, graph-based and knowledge-enhanced approaches offer novel perspectives for enriching

semantic representation and meaning. This survey not only organizes recent advancements in STS and semantic representation literature but also highlights practical applications and delineates future research opportunities within this rapidly evolving field.

# CHAPTER 1

# Foundations Of Semantic Textual Similarity

Semantic Textual Similarity (STS) refers to the task of quantifying the degree of semantic equivalence between two textual units, typically sentences or short texts. It plays a critical role in many natural language processing (NLP) applications, including information retrieval, question answering, text summarization, and machine translation. Unlike syntactic similarity, STS focuses on the meaning conveyed by the text, often requiring deep contextual and semantic understanding.

## 1.1 Motivation and Significance

The emergence of transformer-based models such as BERT, RoBERTa, and ALBERT has dramatically transformed the landscape of semantic understanding in NLP tasks, including semantic textual similarity. Since 2019, the field has seen a surge in robust architectures and optimization strategies that have redefined how sentence representations are learned and compared.

Despite the progress, several challenges remain. Traditional STS approaches were limited by shallow contextual understanding and rigid embedding methods. Transformer models addressed these limitations with self-attention mechanisms and contextual embeddings, but they introduced new issues like high computational costs, difficulty in interpretability, and sensitivity to domain shifts [4, 5, 7, 8].

Recent works such as SimCSE [27], MNet-Sim [3], and CLEAR [17] have proposed efficient and accurate models for sentence similarity, leveraging contrastive learning, graph structures, and residual attention. These innovations emphasize the need to survey and categorize post-2019 models to better understand their contributions, trade-offs, and applicability to real-world scenarios.

Moreover, with the rise of lightweight and quantized models such as DistilBERT [15], TinyBERT [34], and MobileBERT [31], STS is becoming feasible on resource-constrained devices, further broadening its impact. This survey aims to systematically explore the landscape of transformer-based STS models developed after 2019, highlighting advances in architecture design, training strategies, and evaluation methodologies.

## 1.2 Problem Definition

Semantic Textual Similarity (STS) refers to the task of quantitatively assessing the degree of semantic equivalence between a pair of textual units, typically sentences or short paragraphs. Formally, given two text fragments $s_1$ and $s_2$, the objective of an STS system is to compute a similarity score $\text{sim}(s_1, s_2) \in \mathbb{R}$, where higher values indicate greater semantic similarity.

Let $\mathcal{S} = \{(s_1^{(i)}, s_2^{(i)}, y^{(i)})\}_{i=1}^{N}$ denote a dataset of $N$ sentence pairs, where each pair $(s_1^{(i)}, s_2^{(i)})$ is annotated with a human-assigned similarity score $y^{(i)} \in [0, 5]$. The goal is to learn a function $f_\theta : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, parameterized by $\theta$, such that:

$$f_\theta(s_1, s_2) \approx y$$

This similarity estimation must go beyond surface-level lexical overlap and incorporate deep semantic understanding, including contextual usage, syntactic structure, word sense disambiguation, and discourse-level semantics.

The STS task is inherently regression-based when continuous similarity scores are used, although it can also be cast as a classification or ranking problem in specific settings. Performance is typically evaluated using statistical metrics such as Pearson's $r$, Spearman's $\rho$, and mean squared error (MSE), which reflect the correlation and divergence between predicted and actual similarity scores [9, 20, 24].

Due to its foundational nature, the STS problem serves as a benchmark for evaluating semantic representation models and underpins a wide range of downstream applications including duplicate question detection, paraphrase identification, semantic search, and dialogue systems [5, 8, 21].

## 1.3 Evolution from Traditional to Contemporary Approaches

The Semantic Textual Similarity (STS) task has undergone a significant evolution, driven by advances in both theoretical linguistics and computational models. In the early stages, STS approaches predominantly relied on traditional methods based on surface-level features, lexical overlap, and syntactic parsing. These methods, although interpretable and computationally efficient, were often inadequate in capturing deep semantic nuances, especially in the presence of lexical variation, polysemy, and contextual ambiguity.

Traditional methods primarily utilized techniques such as:

- **Lexical similarity measures:** including Jaccard index, Cosine similarity of bag-of-words (BoW) or TF-IDF vectors, and word overlap ratios.

- **Syntactic parsing:** leveraging dependency and constituency parsers to compute structural similarity.

- **Knowledge-based measures:** using resources such as WordNet to compute path-based or information-theoretic similarity between word senses.

While these techniques provided a baseline for semantic similarity, their inability to incorporate contextual information limited their performance on more complex linguistic phenomena. The advent of distributional semantics marked a paradigm shift. Models such as Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), and word embeddings like Word2Vec and GloVe enabled data-driven representations of words in dense vector spaces, capturing semantic relatedness through vector proximity.

The contemporary era of STS has been dominated by neural architectures, particularly those based on deep learning and transformer models. Sentence encoders, such as InferSent, Universal Sentence Encoder (USE), and Sentence-BERT (SBERT), have shown substantial improvements in capturing compositional semantics. Furthermore, pre-trained language models like BERT, RoBERTa, XLNet, and DeBERTa have redefined STS benchmarks by leveraging contextualized embeddings, transfer learning, and fine-tuning strategies.

In recent years, graph-based neural networks, semantic role labeling (SRL), and knowledge graph integration have further enhanced the modeling of semantic relationships by incorporating structural and relational information. These contemporary approaches offer a more holistic understanding of semantic similarity by considering the underlying meaning, context, and syntactic-semantic dependencies between sentences.

This transition from traditional to contemporary approaches reflects an increasing emphasis on deeper, more abstract, and context-sensitive representations, aligning the STS task more closely with human-like understanding of language.

## 1.4 Thesis Structure and Organization

Following the foundational discussion on the evolution of Semantic Textual Similarity (STS) methodologies, this thesis is organized to provide a systematic and in-depth examination of the subject. Chapter 2 presents a comprehensive analysis of semantic similarity approaches, encompassing the underlying transformer architectures, domain-specific frameworks, multimodal integration strategies, graph-theoretic models, and knowledge-enhanced systems. Chapter 3 delves into semantic similarity evaluation resources, including benchmark datasets and the application of Semantic Role Labeling (SRL) for contextual comprehension. Chapter 4 explores innovative applications that extend the utility of STS in diverse domains, showcasing systems such as EvoSimSearch, LLM-DrawNorm, SympLink, and others that highlight the breadth of semantic modeling capabilities. Finally, Chapter 5 synthesizes the key findings, discusses broader research implications, and outlines emerging trends and potential directions for future inquiry in the field of semantic similarity. The thesis concludes with a detailed list of references supporting the theoretical and empirical work presented.

# CHAPTER 2

## Comprehensive Analysis of Semantic Similarity Approaches

In the previous chapter, we explored classical and hybrid similarity functions through a mathematical and statistical lens, highlighting their limitations in capturing contextual and semantic nuances of natural language. With the advent of deep learning, especially Transformer-based architectures, semantic similarity measurement has entered a new era of contextualized understanding. This chapter delves into the architecture, strengths, and evaluation of state-of-the-art Transformer models and their ensembles in semantic similarity tasks.

## 2.1 Transformer Models in Semantic Analysis

Transformer models have significantly advanced the field of semantic similarity by enabling deep contextual understanding of text. These models leverage self-attention mechanisms to weigh the significance of each token relative to others in a sentence, allowing them to capture syntactic and semantic relationships effectively. In this section, we present a comparative analysis of leading models and highlight novel approaches that have recently emerged.

Table 2.1 summarizes performance metrics of several top Transformer-based models on standard semantic similarity benchmarks, measured in Pearson and Spearman correlation.

Table 2.1: Performance of State-of-the-Art Transformer Models on Semantic Similarity Tasks

| **Model** | **Pearson** | **Spearman** | **Reference** |
|---|---|---|---|
| StructBERT/RoBERTa Ensemble | 0.928 | 0.924 | [2, 7] |
| MNet-Sim | 0.927 | 0.931 | [3] |
| T5-11B | 0.925 | 0.921 | [4] |
| ALBERT | 0.925 | - | [5] |
| XLNet (Single) | 0.925 | - | [6] |
| RoBERTa | 0.922 | - | [7] |
| T5-3B | 0.906 | 0.898 | [4] |
| MLM+del-word (CLEAR) | 0.905 | - | [1] |
| RealFormer | 0.9011 | 0.8988 | [11] |

These models are designed to optimize different aspects of language modeling such as token efficiency (Charformer), disentangled attention (DeBERTa), and lightweight architecture (ALBERT). Several of them demonstrate performance nearing the empirical upper bounds on semantic similarity benchmarks.

## 2.2 Transformer-based Methods for Semantic Similarity

The development of deep learning brought transformer based models which revolutionized the process of seman tic similarity assessment. Textual semantics get encoded ef fectively through large-scale pretraining together with self attention mechanisms in these models. Recent transformer based methods introduced after 2021 are the focus of this section.

### FarSSiBERT: Persian Social Media Text Understanding

FarSSiBERT is a transformer tailored for Persian informal texts using a pre-training dataset of over 104M documents. It outperforms multilingual models in both FarSSiM and FarSICK datasets [52].

The following results showed that FarSSiBERT outperforms ParsBERT, laBSE, and multilingual BERT in both datasets, achieving the highest correlation scores:

| **Model** | **Pearson(FarSSiM)** | **Spearman(FarSSiM)** |
|---|---|---|
| FarSSiBERT-104M | 0.770 | 0.643 |
| FarSSiBERT-6M | 0.740 | 0.621 |
| laBSE | 0.725 | 0.643 |
| ParsBERT | 0.704 | 0.624 |
| Multilingual BERT | 0.618 | 0.480 |

### DeBERTa-v3-LSTM Ensemble: AI vs Human Text

Combines DeBERTa-v3 with Bi-LSTM and linear attention pooling for discriminating AI-generated content. Shows improvement in Pearson, F1, and AUC metrics through adversarial fine-tuning and shuffling mechanisms [47].

| **Model** | **Pearson** | **MSE** | **F1-score** | **AUC** |
|---|---|---|---|---|
| DeBERTa-v3-large | 86.1% | 0.015 | 88.5% | 91.2% |
| + Bi-LSTM | 86.6% | 0.014 | 89.1% | 92.3% |
| + Linear Att. Pooling | 86.8% | 0.013 | 89.4% | 92.8% |
| + Target Shuffling | 87.2% | 0.012 | 90.1% | 93.5% |
| Ensemble Model | 87.5% | 0.011 | 91.2% | 94.7% |

The experimental results show that Bi-LSTMs and attention pooling together with target shuffling enhance performance in successive stages.

### BeeFormer: Recommendation-Aware Semantics

BeeFormer enhances similarity for recommender systems using user-item interactions with semantic alignment. Employs a matrix loss minimizing Frobenius norm differences, optimized for cold-start scenarios [49].

### RWKV: Linear Attention Meets Recurrence

RWKV integrates RNN-style recurrence in a linear attention setup, enabling scalability. This model does layer-wise analysis and baseline comparison for semantic similarity. Though performance lags in similarity tasks, it performs more efficient for long sequences[45].

### 3D Siamese Networks: Spatial Semantic Encoding

This model incorporates spatial attention modules and adaptive feature transformations to improve contextual embedding in Siamese networks. Outperforms SBERT and ColBERT across multiple datasets including QQP and SNLI [89].

## 2.3 Contrastive Learning in Semantic Similarity

Contrastive learning approaches, such as CSS, AspectCSE, and PCC-Tuning, have revolutionized the way models learn representations by pulling similar sentences together and pushing dissimilar ones apart. Sentence embeddings together with semantic similarity evaluation make its learning process via contrastive learning the current mainstream approach. By implementing this technique representation quality improves because semantically related texts receive closer spatial mapping in the embedding space, along with dissimilar texts distant from each other.

### Contrastive Semantic Similarity(CSS)

The CSS (Contrastive Semantic Similarity) framework [60] introduces a novel approach to quantifying uncertainty in responses generated by Large Language Models (LLMs). Unlike traditional Natural Language Inference (NLI) methods, which rely on class probabilities, CSS uses a CLIP-based contrastive feature extraction to better represent semantic relationships. The method generates text embeddings and computes semantic similarity using the Hadamard product, as shown in Equation 2.1. Further, the Graph Laplacian is used to enhance clustering and uncertainty estimation by analyzing the eigenvalues of the Laplacian matrix, which correlate with the number of semantic clusters.

$$CSS_{i,j} = r_i \odot r_j \tag{2.1}$$

### AspectCSE: Aspect-based Semantic Similarity using Contrastive Learning

[AspectCSE, [90]] is a contrastive learning approach that improves sentence embeddings by focusing on specific aspects of meaning. By incorporating structured knowledge from the Wikidata graph, AspectCSE learns multi-aspect sentence embeddings that help in analyzing the semantic similarity of texts across multiple dimensions. The optimization is carried out using a cosine similarity-based loss function, which links positive samples (same-aspect sentences) and negative samples (different-aspect sentences), as outlined in Equation 2.2. This method improves retrieval performance by 3.97% compared to existing models, as it benefits from aspect-specific semantic relations in structured knowledge sources.

$$\ell_i = -\log \frac{\exp(\mathrm{sim}(h_{a_i}, h_{a_i^+})/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(h_{a_i}, h_{a_j^+})/\tau) + \exp(\mathrm{sim}(h_{a_i}, h_{a_j^-})/\tau)} \tag{2.2}$$

### PCC-Tuning: Breaking the Ceiling in Contrastive Learning for Text Similarity

[PCC-Tuning, [55]] enhances contrastive learning for Semantic Textual Similarity (STS) tasks by incorporating Pearson's correlation coefficient into the loss function. This modification overcomes the limitations of traditional InfoNCE loss functions, achieving higher performance in text similarity tasks. Through a two-step training process involving large-scale NLI dataset contrastive learning followed by fine-tuning with Pearson's correlation loss, PCC-Tuning reaches a Spearman correlation of 87.86, surpassing the previous performance limit of 87.5. This advancement is critical for precise semantic matching in text pairs.

$$\ell_p = -r + 1 \quad \text{where} \quad r \in [0, 2] \tag{2.3}$$

## 2.4 Domain Specific Semantic Similarity

The successful interpretation of specialized domains heavily depends on semantic similarity evaluation in NLP applications because standard NLP methods do not effectively identify complex domain relationships. Multiple research studies have developed specialized semantic similarity models which focus on financial healthcare recruitment and other domain fields.

### CXR-BERT: Semantic Similarity in Chest X-ray Reports

[CXR-BERT, [73]] enhances chest X-ray report generation through a contrastive representation learning framework. The model uses cosine similarity to compare the embeddings of chest X-ray reports generated by CXR-BERT. A semantic similarity reward function is introduced, which adjusts the influence of the similarity measure during optimization:

$$S(Rg, Rr) = \frac{E(Rg) \cdot E(Rr)}{\|E(Rg)\|\|E(Rr)\|}$$

A scaling parameter $\lambda$ is used for model optimization, with experimental results on MIMIC-CXR and Open-i IU X-ray datasets demonstrating improved performance over traditional models.

## Financial-STS: Detecting Subtle Semantic Shifts in Financial Narratives

[Financial-STS, [67]] identifies semantic shifts in financial narratives by using Large Language Models (LLMs). The model learns a function $\Phi$ that maps pairs of financial statements to a similarity score, where a lower score indicates a greater semantic shift. The network minimizes a triplet loss:

$$\max(\cos(si, ni) - \cos(si, pi) + \epsilon, 0)$$

Experimental results show that Financial-STS outperforms traditional models, achieving higher AUC metrics and effectively detecting intensified sentiment and emerging situations.

## ICD-STS: Enhancing ICD-based Similarity

[ICD-STS, [83]] improves the similarity assessment of ICD codes by introducing a scaling term to adjust for comorbidity variations:

$$ST(A, B) = \frac{\mathrm{SetSim}(A, B)}{\min(|A|, |B|) + \log(1 + |A| - |B|)}$$

The method improves expert-annotated similarity score correlations and enhances patient similarity assessment for precision medicine applications.

### PatentBERT: Patent Document Matching Using Ensemble BERT

[PatentBERT, [75]] improves semantic similarity evaluations in patent documents using a GPT-4-based framework. The model performs label generation for patent documents and compares them based on semantic understanding, surpassing traditional methods like ROUGE and BLEU in matching expert-validated similarity standards.

### VacancySBERT: Vacancy Matching with Transformer-based Models

[VacancySBERT, [86]] matches recruitment job titles with skills using a Siamese network architecture based on Sentence-BERT (SBERT). The model optimizes text similarity operations and utilizes a Multiple Negative Ranking (MNR) loss to improve skill-based matching, achieving better results compared to models like JobBERT.

### BM$\chi$: Information Retrieval Enhancement

[BM$\chi$, [50]] improves BM25's performance in information retrieval by incorporating entropy-weighted similarity and Weighted Query Augmentation (WQA). The modified scoring function enhances semantic understanding and retrieval accuracy:

$$\text{score}(D, Q) = \sum_i \text{IDF}(qi)\cdot F(qi, D)\cdot(\alpha+1)\left[\frac{F(qi, D)}{|D|} + \alpha \cdot E + \beta \cdot E(qi) \cdot S(Q, D)\right]$$

Experimental results indicate BM$\chi$ outperforms BM25 across various retrieval benchmarks, demonstrating effective results in both short and long-context information retrieval tasks.

## 2.5 Multi-modal Semantic Representation and Comparison

The Multi-modal Semantic Similarity approach deals with measuring correlation across multiple modes which include text as well as image and audio types. The similarity estimation process gets enhanced through the use of multiple data sources when traditional textual embeddings methods become insufficient. This part explores modern developments in multi-modal semantic similarity which it organizes according to fundamental techniques along with their resulting contributions.

### DuSSS: Vision-Language Models for Medical Image Segmentation

The Dual Contrastive Learning (DCL) framework optimizes vision-language models (VLMs) for semi-supervised medical image segmentation by enhancing image-text semantic alignment. It incorporates Cross-Modal Contrastive Learning (CMC) and Intra-Modal Contrastive Learning (IMC) to achieve better alignment. The CMC function utilizes an uncertainty-constrained cosine similarity function to drive matching elements together and push mismatched ones apart. The objective is optimized using InfoNCE loss, ensuring that positive image-text pairs are closer in embedding space. DCL has demonstrated improvements in segmentation performance, achieving an 82.52% Dice score compared to state-of-the-art methods [48].

### TexIm-FAST: Text-to-Image Representations for Similarity Evaluation

[TexIm-FAST, [57]] addresses the computational challenges of traditional text embeddings by converting text data into grayscale image representations using CNN-based Variational Autoencoders (VAEs). The method applies self-supervised learning and quantization to optimize feature learning. Its approach results in reduced system memory requirements while achieving a 6% improvement over traditional methods in tasks like semantic textual similarity (STS) evaluation.

### CSFNet: Real-time RGB-X Segmentation using Cosine Similarity Fusion

[CSFNet, [53]] employs a dual-branch encoder architecture for RGB and auxiliary modality inputs, optimizing feature fusion using cosine similarity. The network uses an attention-based fusion mechanism to integrate cross-modal features, significantly improving segmentation performance with reduced computational expense. Experimental results show that CSFNet outperforms existing RGB-D models in real-time segmentation, achieving a mean Intersection over Union (mIoU) of 76.36% on the Cityscapes dataset and 91.40% on the ZJU dataset.

### SeSS: Semantic Similarity Scoring for Multi-modal Data

[SeSS, [58]] is a new metric for evaluating visual image similarity based on Scene Graph Generation (SGG). It computes semantic similarity between images using object-relation graphs, which better aligns with human visual perception. Experimental results show that SeSS outperforms traditional metrics, providing more accurate similarity assessments for compressed or noisy images.

### ImageGen-SSC: Measuring Image-Generative Semantic Communication

Image-Generative Semantic Communication enhances image transmission by extracting and transmitting semantic features instead of full images, thus reducing transmission costs. This method computes a semantic similarity score between the original and generated images using both textual similarity (via BERTScore) and segmentation accuracy (via Segmentation Matching Rate). Experimental verification shows that the technique reduces data transmission by 14 times compared to JPEG while maintaining image quality [63].

## 2.6 Graph-Theoretic Approaches for Semantic Relationship Modeling

GraphSQLSim, RDF-RecSys, ReMatch, SemDiff, PEM, and DGNN-SRL represent significant advances in semantic relationship modeling using graph-based approaches. These techniques utilize graphs to capture semantic structures in various contexts such as SQL query assessment, recommendation systems, knowledge graph matching, binary similarity detection, and semantic textual similarity evaluation.

### GraphSQLSim: SQL Query Similarity Using Graphs

[GraphSQLSim, [66]] models SQL query similarity by transforming queries into implicit network nodes and using weighted edit operations to find the most cost-effective transformation. Given an initial query $Q_s$ and reference query $Q_r$, the semantic distance is computed as:

$$d(Q_s, Q_r) = \sum_{i=1}^{n} c(e_i)$$

where $e_i$ is an individual edit operation in the shortest path sequence, and $c(e_i)$ is its assigned cost. The grading score is derived as:

$$\text{score} = \max(0, P_{\max} - \alpha d(Q_s, Q_r))$$

where $P_{\max}$ is the maximum score and $\alpha$ is a scaling factor.

Experimental results show that this approach approximates human grading, outperforming traditional methods in fairness and comprehensibility.

## RDF-RecSys: RDF Graphs for Recommender Systems

[RDF-RecSys, [87]] enhances recommendation systems by combining text and numeric data through topic-based contextual information. Similarity between two RDF triplets $a_1 = \langle s_1, p_1, o_1 \rangle$ and $a_2 = \langle s_2, p_2, o_2 \rangle$ is defined as:

$$\text{Sim}(a_1, a_2) = \frac{1}{N} \sum_{i \in P} \omega_i \text{Sim}_1(a_{i1}, a_{i2}) + \gamma \text{Sim}_2(a_{o1}, a_{o2})$$

where $P$ represents the components of the triplet, and $\omega_i$, $\gamma$ are weights assigned to subject, predicate, and object components.

The method showed a significant improvement in recommendation accuracy, surpassing Jaccard similarity and TF-IDF models in 82.4

## ReMatch: Efficient Knowledge Graph Matching

[ReMatch, [64]] improves knowledge graph alignment using semantic motif matching and structural similarity. The similarity between two AMR graphs $G_1$ and $G_2$ is calculated using Jaccard similarity:

$$\text{Rematch}(G_1, G_2) = \frac{|M(G_1) \cap M(G_2)|}{|M(G_1) \cup M(G_2)|}$$

Structural similarity is assessed using the Randomized AMRs with Rewired Edges (RARE) benchmark:

$$\text{Sim}(G, G') = 1 - \frac{|E' - E|}{|E|}$$

where $E$ and $E'$ are the edge sets of the original and perturbed graphs. ReMatch outperforms existing AMR similarity metrics, improving efficiency and accuracy.

### SemDiff: Binary Similarity Detection with Key-Semantics Graphs

[SemDiff, [84]] uses key-semantics graphs (KSG) to detect matching binary functions. The similarity between two binary functions $G_1$ and $G_2$ is calculated using Locality-Sensitive Hashing (LSH):

$$\text{Sim}(G_1, G_2) = \frac{|LSH(G_1) \cap LSH(G_2)|}{|LSH(G_1) \cup LSH(G_2)|}$$

This method provides better results than traditional tools for detecting cross-compiler and obfuscation variations.

### PEM: Probabilistic Execution Models for Binary Similarity

[PEM], [80] models binary semantics through statistical analysis of program code flow. The binary program $P$ is represented as a distribution of input-output behaviors:

$$D_P = \{(x, O_V(P(x)))|x \in X\}$$

where $x$ is an input and $O_V(P(x))$ denotes externally observable values. The system uses path sampling to match execution pathways between different executable programs. Experimental results demonstrate that PEM achieves 96

### DGNN-SRL: Deep Graph Neural Networks with SRL Graphs

[DGNN-SRL, [92]] enhances Semantic Textual Similarity (STS) by integrating Semantic Role Labeling (SRL) graphs with Deep Graph Neural Networks (DGNN). The reconstruction loss used in DGNN is defined as:

$$\text{loss}_{\text{reconstruction}} = \text{SmoothL1Loss}(\text{output}, \text{input})$$

The evaluation results from the STS2017 and SICK datasets show that integrating SRL graphs with DGNN improves performance, with the SRL+SDG model outperforming standard DG graphs in sentence similarity evaluation.

| System | Pearson | Spearman |
|---|---|---|
| SRL+SDG | 0.9267 | 0.9253 |
| SRL+DG | 0.9272 | 0.9261 |
| DG | 0.9275 | 0.9264 |

This approach outperforms standard transformer models, especially with lengthy sentences, and shows notable improvements when applied to the RoBERTa transformer.

# CHAPTER 3

## 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 [3, 27, 4].

## 3.1 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. [4] | 3,108 papers, 25 benchmarks |
| MRPC | Microsoft Research Paraphrase Corpus with 5,801 sentence pairs labeled as paraphrases or not [27] | 768 papers, 5 benchmarks |
| SICK | Sentences Involving Compositional Knowledge annotated for relatedness and entailment [3] | 342 papers, 5 benchmarks |
| SentEval | Toolkit for evaluating universal sentence encoders across multiple tasks including STS [27] | 166 papers, 2 benchmarks |
| MTEB | Massive Text Embedding Benchmark with 56 datasets covering 8 tasks in 112 languages [25] | 133 papers, 8 benchmarks |
| CARER | Contextualized Affect Representations for Emotion Recognition with noisy distant-supervised annotations [27] | 119 papers, 1 benchmark |
| STS Benchmark | Dataset from STS tasks at SemEval (2012–2017), including image captions and forum texts [27] | 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 [27] | 22 papers, 1 benchmark |
| CxC | Crisscrossed Captions dataset with 247k+ human annotations on images and captions [27] | 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 [27] | 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 [3] | 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 3.1: Overview of Datasets for Semantic Similarity Evaluation

# CHAPTER 4

# Benchmarking STS Models

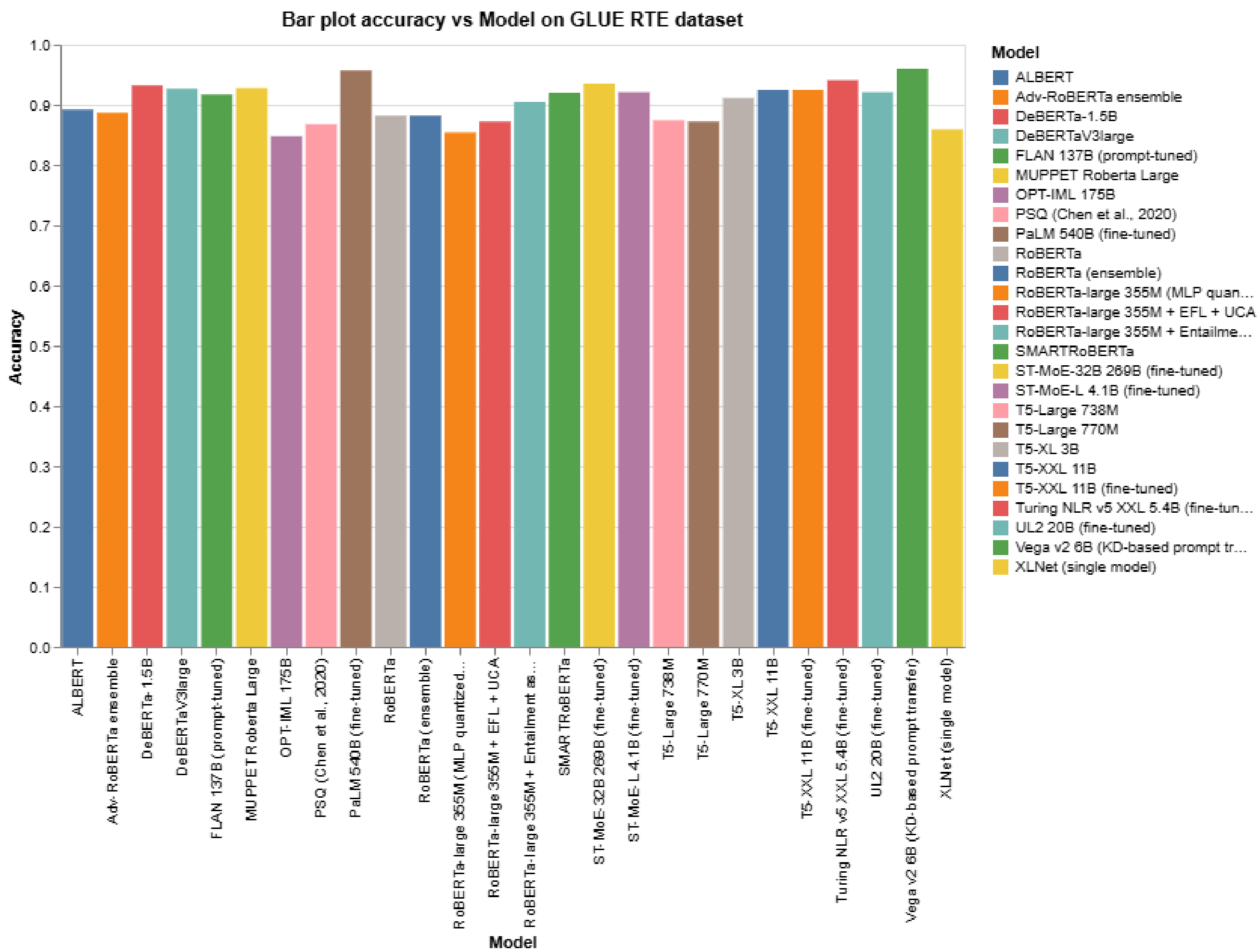
Semantic Textual Similarity (STS) is a crucial task in natural language processing that evaluates the semantic similarity between sentence pairs. STS models have evolved significantly over the years, from traditional lexical approaches to advanced transformer-based models. This chapter aims to provide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

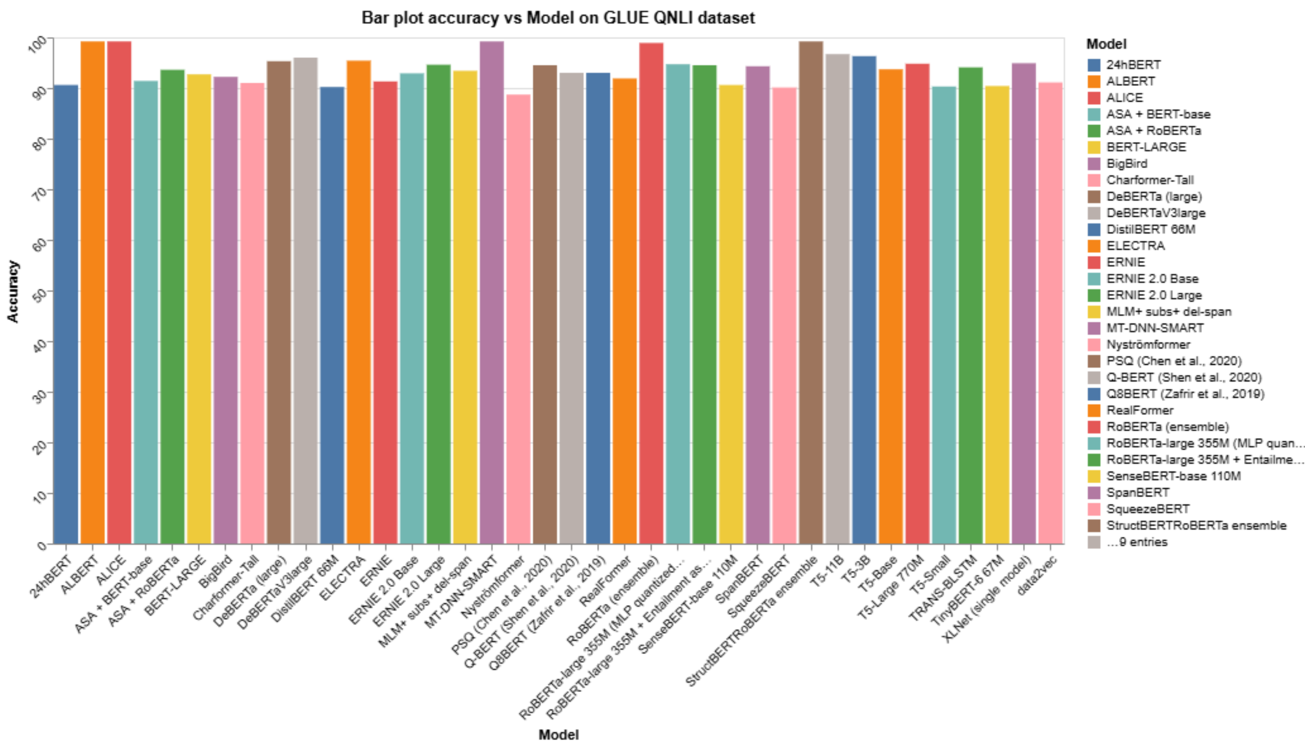
## 4.1 Datasets and Metrics

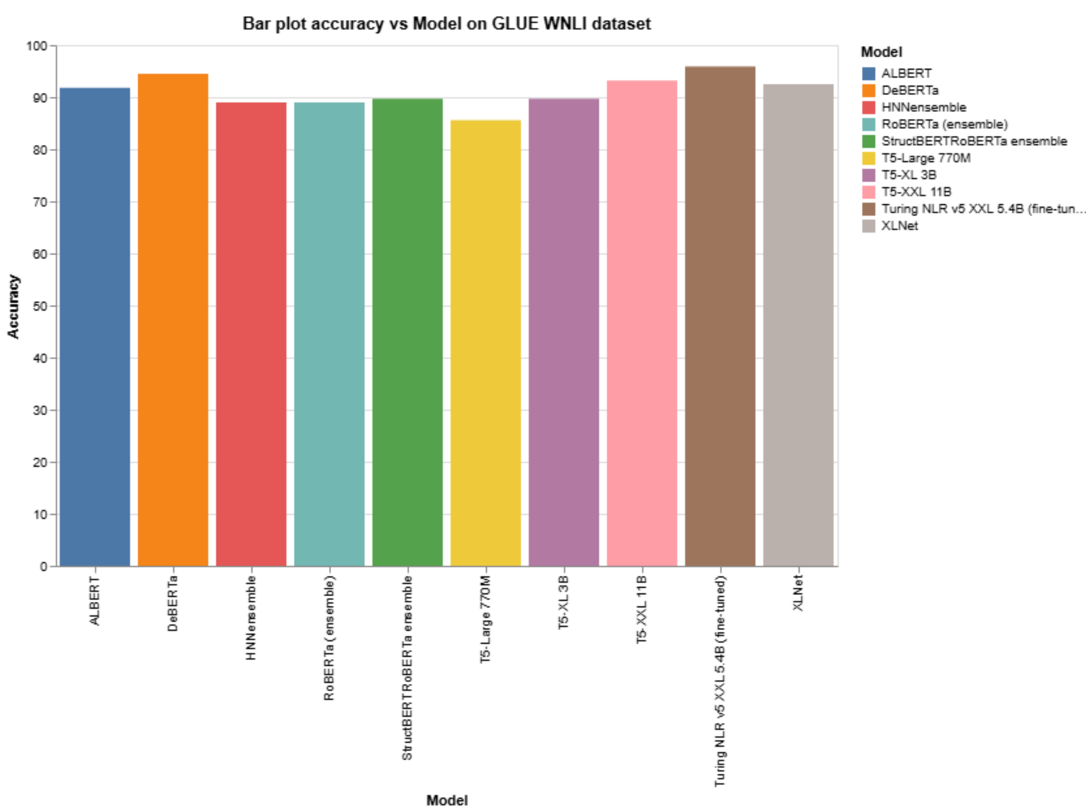
The objective of benchmarking STS models is to assess their effectiveness in capturing semantic similarities between sentences. Recent advancements in pre-trained language models, such as BERT, RoBERTa, and T5, have shown remarkable performance in various natural language processing tasks, including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

## 4.2 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.




Bar plot accuracy vs Model on GLUE QNLI dataset
Accuracy
Model
0
10
20
30
40
50
60
70
80
90
100
24hBERT
ALBERT
ALICE
ASA + BERT-base
ASA + RoBERTa
BERT-LARGE
BigBird
Charformer-Tall
DeBERTa (large)
DeBERTaV3large
DistilBERT 66M
ELECTRA
ERNIE
ERNIE 2.0 Base
ERNIE 2.0 Large
MLM+ subs+ del-span
MT-DNN-SMART
Nyströmformer
PSQ (Chen et al., 2020)
Q-BERT (Shen et al., 2020)
Q8BERT (Zafrir et al., 2019)
RealFormer
RoBERTa (ensemble)
RoBERTa-large 355M (MLP quantized…
RoBERTa-large 355M + Entailment as…
SenseBERT-base 110M
SpanBERT
SqueezeBERT
StructBERTRoBERTa ensemble
T5-11B
T5-3B
T5-Base
T5-Large 770M
T5-Small
TRANS-BLSTM
TinyBERT-6 67M
XLNet (single model)
data2vec
Model
24hBERT
ALBERT
ALICE
ASA + BERT-base
ASA + RoBERTa
BERT-LARGE
BigBird
Charformer-Tall
DeBERTa (large)
DeBERTaV3large
DistilBERT 66M
ELECTRA
ERNIE
ERNIE 2.0 Base
ERNIE 2.0 Large
MLM+ subs+ del-span
MT-DNN-SMART
Nyströmformer
PSQ (Chen et al., 2020)
Q-BERT (Shen et al., 2020)
Q8BERT (Zafrir et al., 2019)
RealFormer
RoBERTa (ensemble)
RoBERTa-large 355M (MLP quan…
RoBERTa-large 355M + Entailme…
SenseBERT-base 110M
SpanBERT
SqueezeBERT
StructBERTRoBERTa ensemble
…9 entries


Bar plot accuracy vs Model on GLUE WNLI dataset
Accuracy
Model
0
10
20
30
40
50
60
70
80
90
100
ALBERT
DeBERTa
HNNensemble
RoBERTa (ensemble)
StrucBERTRoBERTa ensemble
T5-Large 770M
T5-XL 3B
T5-XXL 11B
Turing NLR v5 XXL 5.4B (fine-tuned)
XLNet
Model
ALBERT
DeBERTa
HNNensemble
RoBERTa (ensemble)
StructBERTRoBERTa ensemble
T5-Large 770M
T5-XL 3B
T5-XXL 11B
Turing NLR v5 XXL 5.4B (fine-tun…
XLNet

| Rank | Model | P. Corr | Paper | Year | Tags |
|---|---|---|---|---|---|
| 1 | MT-DNN-SMART | 0.929 | SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization [1] | 2019 | |
| 2 | StructBERT / RoBERTa ensemble | 0.928 | StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding [2] | 2019 | Transformer |
| 3 | Mnet-Sim | 0.927 | MNet-Sim: A Multi-layered Semantic Similarity Network to Evaluate Sentence Similarity [3] | 2021 | |
| 4 | T5-11B | 0.925 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [4] | 2019 | Transformer |
| 5 | ALBERT | 0.925 | ALBERT: A Lite BERT for Self-supervised Learning of Language Representations [5] | 2019 | Transformer |
| 6 | XLNet (single model) | 0.925 | XLNet: Generalized Autoregressive Pretraining for Language Understanding [6] | 2019 | Transformer |
| 7 | RoBERTa | 0.922 | RoBERTa: A Robustly Optimized BERT Pretraining Approach [7] | 2019 | Transformer |
| 8 | ELECTRA | 0.921 | ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators [8] | 2020 | |
| 9 | RoBERTa-large 355M (MLP quantized, fine-tuned) | 0.919 | LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale [9] | 2022 | |
| 10 | PSQ (Chen et al., 2020) | 0.919 | A Statistical Framework for Low-bitwidth Training of Deep Neural Networks [10] | 2020 | |

*Continued on next page*

| Rank | Model | P. Corr | Paper | Year | Tags |
|---|---|---|---|---|---|
| 11 | RoBERTa-large 355M + Entailment as Few-shot | 0.918 | Entailment as Few-Shot Learner [11] | 2021 | |
| 12 | ERNIE 2.0 Large | 0.912 | ERNIE 2.0: A Continual Pre-training Framework for Language Understanding [12] | 2019 | |
| 13 | Q-BERT (Shen et al., 2020) | 0.911 | Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT [13] | 2019 | |
| 14 | Q8BERT (Zafrir et al., 2019) | 0.911 | Q8BERT: Quantized 8Bit BERT [14] | 2019 | |
| 15 | ELECTRA (no tricks) | 0.910 | ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators [8] | 2020 | |
| 16 | DistilBERT 66M | 0.907 | DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [15] | 2019 | |
| 17 | T5-3B | 0.906 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [4] | 2019 | Transformer |
| 18 | MLM+ del-word | 0.905 | CLEAR: Contrastive Learning for Sentence Representation [17] | 2020 | |
| 19 | RealFormer | 0.9011 | RealFormer: Transformer Likes Residual Attention [18] | 2020 | Transformer |
| 20 | T5-Large | 0.899 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [4] | 2019 | Transformer |

*Continued on next page*

| Rank | Model | P. Corr | Paper | Year | Tags |
|---|---|---|---|---|---|
| 21 | SpanBERT | 0.899 | SpanBERT: Improving Pre-training by Representing and Predicting Spans [19] | 2019 | Transformer |
| 22 | T5-Base | 0.894 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [4] | 2019 | Transformer |
| 23 | ERNIE 2.0 Base | 0.876 | ERNIE 2.0: A Continual Pre-training Framework for Language Understanding [12] | 2019 | Transformer |
| 24 | Charformer-Tall | 0.873 | Charformer: Fast Character Transformers via Gradient-based Subword Tokenization [20] | 2021 | |
| 25 | T5-Small | 0.856 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [4] | 2019 | Transformer |
| 26 | ERNIE | 0.832 | ERNIE: Enhanced Language Representation with Informative Entities [21] | 2019 | Transformer |
| 27 | 24hBERT | 0.820 | How to Train BERT with an Academic Budget [22] | 2021 | Transformer |
| 30 | AnglE-LLaMA-13B | 0.8969 | AnglE-optimized Text Embeddings [23] | 2023 | |
| 31 | ASA + RoBERTa | 0.892 | Adversarial Self-Attention for Language Understanding [24] | 2022 | |
| 32 | PromptEOL+CSE +LLaMA-30B | 0.8914 | Scaling Sentence Embeddings with Large Language Models [25] | 2023 | |
| 33 | AnglE-LLaMA-7B | 0.8897 | AnglE-optimized Text Embeddings [23] | 2023 | |

*Continued on next page*

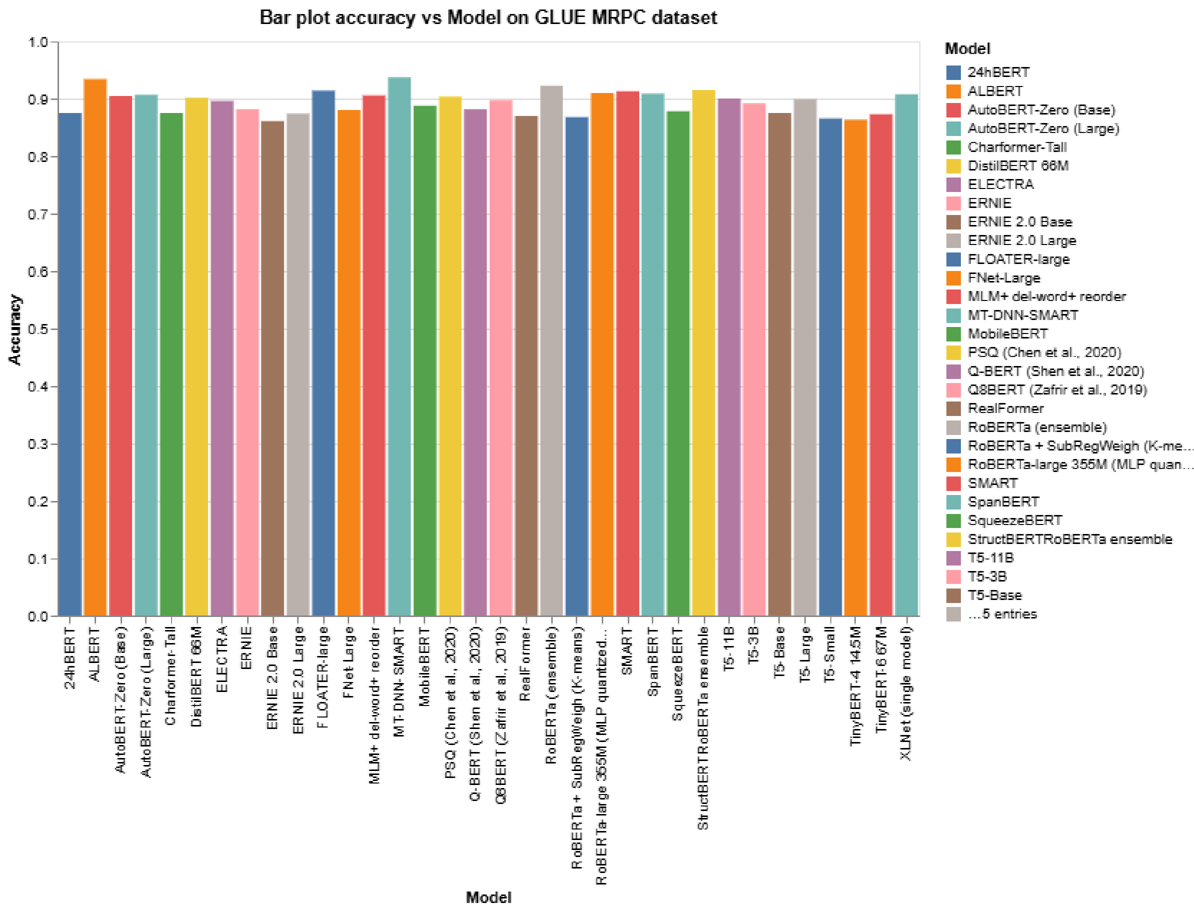
| Rank | Model | P. Corr | Paper | Year | Tags |
|---|---|---|---|---|---|
| 34 | AnglE-LLaMA-7B-v2 | 0.8897 | AnglE-optimized Text Embeddings [23] | 2023 | |
| 35 | T5-Large 770M | 0.886 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [4] | 2019 | |
| 36 | Prompt EOL+CSE +OPT-13B | 0.8856 | Scaling Sentence Embeddings with Large Language Models [25] | 2023 | |
| 37 | Prompt EOL+CSE +OPT-2.7B | 0.8833 | Scaling Sentence Embeddings with Large Language Models [25] | 2023 | |
| 38 | PromCSE-RoBERTa-large (0.355B) | 0.8787 | Improved Universal Sentence Embeddings with Prompt-based Contrastive Learning and Energy-based Learning [29] | 2022 | |
| 39 | BigBird | 0.878 | Big Bird: Transformers for Longer Sequences [26] | 2020 | Transformer |
| 40 | SimCSE-RoBERTa-large | 0.867 | SimCSE: Simple Contrastive Learning of Sentence Embeddings [27] | 2021 | |
| 41 | Trans-Encoder-RoBERTa-large-cross (unsup.) | 0.867 | Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations [28] | 2021 | |

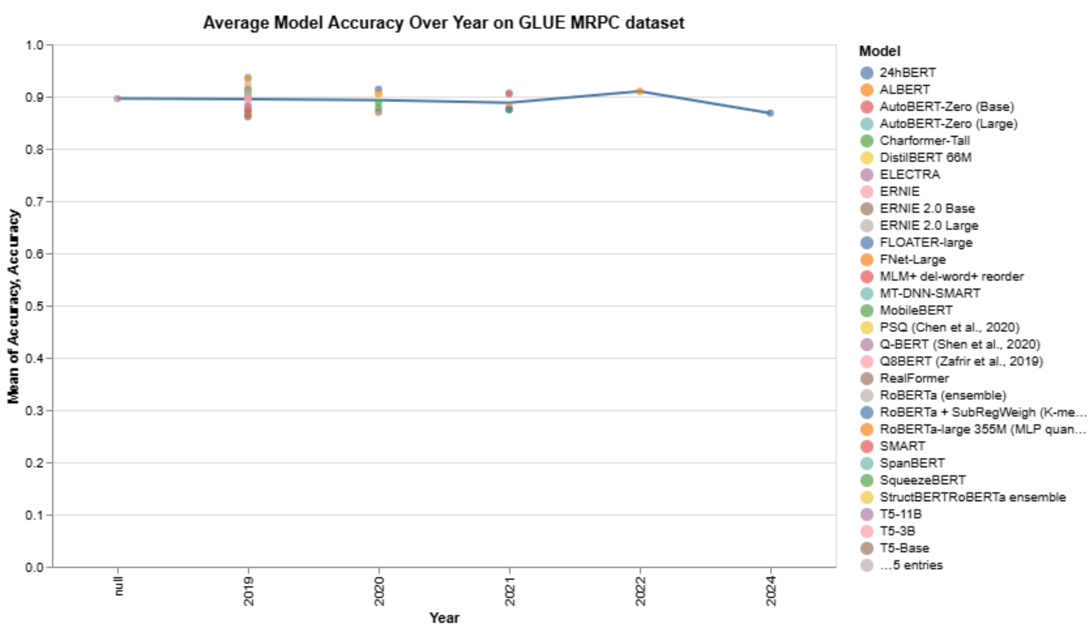
*Continued on next page*

| Rank | Model | P. Corr | Paper | Year | Tags |
|---|---|---|---|---|---|
| 42 | Trans-Encoder-RoBERTa-large-bi (unsup.) | 0.8655 | Trans-Encoder: Unsupervised sentence-pair modelling through self- and mutual-distillations [28] | 2021 | |

## 4.3 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
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

| Rank | Model | Accuracy | F1 | Paper | Year |
|---|---|---|---|---|---|
| 1 | MT-DNN-SMART | 93.7% | 91.7 | SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization [1] | 2019 |
| 2 | ALBERT | 93.4% | | ALBERT: A Lite BERT for Self-supervised Learning of Language Representations [5] | 2019 |
| 3 | RoBERTa (ensemble) | 92.3% | | RoBERTa: A Robustly Optimized BERT Pretraining Approach [7] | 2019 |
| 4 | StructBERT/RoBERTa ensemble | 91.5% | 93.6 | StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding [2] | 2019 |
| 5 | FLOATER-large | 91.4% | | Learning to Encode Position for Transformer with Continuous Dynamical Model [29] | 2020 |
| 6 | SMART | 91.3% | | SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization [1] | 2019 |
| 7 | RoBERTa-large 355M (MLP quantized vector-wise, fine-tuned) | 91.0% | | LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale [9] | 2022 |
| 8 | SpanBERT | 90.9% | | SpanBERT: Improving Pre-training by Representing and Predicting Spans [19] | 2019 |

*Continued on next page*

| Rank | Model | Accuracy | F1 | Paper | Year |
|---|---|---|---|---|---|
| 9 | XLNet (single model) | 90.8% | | XLNet: Generalized Autoregressive Pretraining for Language Understanding [6] | 2019 |
| 10 | AutoBERT-Zero (Large) | 90.7% | | AutoBERT-Zero: Evolving BERT Backbone from Scratch [30] | 2021 |
| 11 | MLM+ del-word+ reorder | 90.6% | | CLEAR: Contrastive Learning for Sentence Representation [17] | 2020 |
| 12 | AutoBERT-Zero (Base) | 90.5% | | AutoBERT-Zero: Evolving BERT Backbone from Scratch [30] | 2021 |
| 13 | PSQ (Chen et al., 2020) | 90.4% | | A Statistical Framework for Low-bitwidth Training of Deep Neural Networks [10] | 2020 |
| 14 | DistilBERT 66M | 90.2% | | DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [15] | 2019 |
| 15 | T5-11B | 90.0% | 91.9 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [16] | 2019 |
| 16 | T5-Large | 89.9% | 92.4 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [16] | 2019 |
| 17 | Q8BERT (Zafrir et al., 2019) | 89.7% | | Q8BERT: Quantized 8Bit BERT [14] | 2019 |
| 18 | ELECTRA | 89.6% | | ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators [8] | 2020 |

*Continued on next page*

| Rank | Model | Accuracy | F1 | Paper | Year |
|---|---|---|---|---|---|
| 19 | T5-3B | 89.2% | 92.5 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [16] | 2019 |
| 20 | MobileBERT | 88.8% | | MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices [31] | 2020 |
| 21 | ERNIE | 88.2% | | ERNIE: Enhanced Language Representation with Informative Entities [21] | 2019 |
| 22 | Q-BERT (Shen et al., 2020) | 88.2% | | Q-BERT: Hessian Based Ultra Low Precision Quantization of BERT [13] | 2020 |
| 23 | FNet-Large | 88% | | FNet: Mixing Tokens with Fourier Transforms [32] | 2021 |
| 24 | SqueezeBERT | 87.8% | | SqueezeBERT: What can computer vision teach NLP about efficient neural networks? [33] | 2020 |
| 25 | Charformer-Tall | 87.5% | 91.4 | Charformer: Fast Character Transformers via Gradient-based Subword Tokenization [20] | 2021 |
| 26 | T5-Base | 87.5% | 90.7 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [16] | 2019 |
| 27 | 24hBERT | 87.5% | | How to Train BERT with an Academic Budget [22] | 2021 |

*Continued on next page*

| Rank | Model | Accuracy | F1 | Paper | Year |
|---|---|---|---|---|---|
| 28 | ERNIE 2.0 Large | 87.4% | | ERNIE 2.0: A Continual Pre-training Framework for Language Understanding [12] | 2019 |
| 29 | TinyBERT-6 67M | 87.3% | | TinyBERT: Distilling BERT for Natural Language Understanding [34] | 2019 |
| 30 | RealFormer | 87.01% | 90.91 | RealFormer: Transformer Likes Residual Attention [18] | 2020 |
| 31 | RoBERTa + SubRegWeigh (K-means) | 86.82% | | SubRegWeigh: Effective and Efficient Annotation Weighing with Subword Regularization [35] | 2024 |
| 32 | T5-Small | 86.6% | 89.7 | Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [16] | 2019 |
| 33 | TinyBERT-4 14.5M | 86.4% | | TinyBERT: Distilling BERT for Natural Language Understanding [34] | 2019 |
| 34 | ERNIE 2.0 Base | 86.1% | | ERNIE 2.0: A Continual Pre-training Framework for Language Understanding [12] | 2019 |

## 4.4 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 [6] | 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 [36] | 2019 |
| 3 | Snorkel MeTaL (ensemble) | 91.5 / 88.5 | 90.1 / 89.7* | Training Complex Models with Multi-Task Weak Supervision [37] | 2018 |
| 4 | TF-KLD | 80.4 / 85.9 | - | Discriminative Improvements to Distributional Sentence Similarity [38] | 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_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 |
| +CoDiaBERT | 88.40 | 82.96 | 88.25 | 79.42 | 53.46 | 82.77 |
| CSRL (Xu et al., 2021) | 88.46 | 81.94 | 89.46 | 78.77 | 51.01 | 82.48 |
| 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 |
| UE2E (Li et al., 2019) | 87.46 | 81.45 | 89.75 | 78.35 | 51.65 | 82.37 |
| LISA (Strubell et al., 2018) | 89.57 | 83.48 | 91.02 | 80.43 | 53.81 | 85.04 |
| SynGCN (Marcheggiani and Titov, 2017) | 90.12 | 84.06 | 91.53 | 82.04 | 54.12 | 85.35 |
| +CoDiaBERT | 91.34 | 86.72 | 91.86 | 82.86 | 56.75 | 85.98 |

*Continued on next page*

| Conversation SRL (Continued) | | | | | | |
|---|---|---|---|---|---|---|
| **Method** | **DuConv** | | | **NewsDialog** | | |
| | **F1_all** | **F1_cross** | **F1_intro** | **F1_all** | **F1_cross** | **F1_intro** |
| POLar (Fei et al., 2022) | 92.06 | 90.75 | 92.64 | 83.45 | 60.68 | 87.96 |
| +CoDiaBERT | 93.72 | 92.86 | 93.92 | 85.10 | 63.85 | 88.23 |

# CHAPTER 5

# Conclusion

The field of Semantic Textual Similarity (STS) has evolved significantly since 2021, driven by innovations in transformer architectures, contrastive learning, and domain-specific applications. This survey has examined these developments across six key areas: transformer-based models, contrastive learning approaches, domain-specific models, multi-modal models, graph-based approaches, and knowledge-enhanced models.

Transformer-based models like FarSSiBERT and DeBERTa-v3 have established new benchmarks in STS tasks, demonstrating superior performance in capturing contextual semantics across various languages and domains. Contrastive learning has emerged as a powerful paradigm for enhancing semantic similarity assessment, with methods like AspectCSE and PCC-Tuning breaking through previous performance ceilings.

Domain-specific models have addressed the unique challenges of specialized fields, with models like CXR-BERT for medical text and Financial-STS for financial narratives showing significant improvements over general-purpose models. Multi-modal semantic similarity has expanded traditional text-based approaches by incorporating visual and audio elements, with models like DuSSS and CLAP demonstrating enhanced semantic representation.

Graph-based approaches have provided a structural dimension to semantic similarity by capturing relationships between textual elements, while knowledge-enhanced models have integrated external structured resources to improve semantic understanding, particularly in domain-specific contexts.

Looking forward, promising research directions include integrating large language models with traditional embedding approaches, developing more efficient models for resource-constrained environments, exploring cross-modal semantic similarity, and addressing challenges of fairness and interpretability.

As STS models become more sophisticated and domain-aware, they will continue to play an increasingly central role in natural language understanding and human-computer interaction systems.

# Bibliography

[1] Jiang, Haoming, et al. ”Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization.” arXiv preprint arXiv:1911.03437 (2019).

[2] Wang, Wei, et al. ”Structbert: Incorporating language structures into pre-training for deep language understanding.” arXiv preprint arXiv:1908.04577 (2019).

[3] Jeyaraj, Manuela Nayantara, and Dharshana Kasthurirathna. ”MNet-Sim: A Multi-layered Semantic Similarity Network to Evaluate Sentence Similarity.” arXiv preprint arXiv:2111.05412 (2021).

[4] Raffel, Colin, et al. ”Exploring the limits of transfer learning with a unified text-to-text transformer.” Journal of machine learning research 21.140 (2020): 1-67.

[5] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942.

[6] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. Advances in neural information processing systems, 32.

[7] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[8] Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

[9] Dettmers, Tim, et al. "Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale." Advances in neural information processing systems 35 (2022): 30318-30332.

[10] Chen, J., Gai, Y., Yao, Z., Mahoney, M. W., & Gonzalez, J. E. (2020). A statistical framework for low-bitwidth training of deep neural networks. Advances in neural information processing systems, 33, 883-894.

[11] Wang, S., Fang, H., Khabsa, M., Mao, H., & Ma, H. (2021). Entailment as few-shot learner. arXiv preprint arXiv:2104.14690.

[12] Sun, Y., Wang, S., Li, Y., Feng, S., Tian, H., Wu, H., & Wang, H. (2020, April). Ernie 2.0: A continual pre-training framework for language understanding. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 05, pp. 8968-8975).

[13] Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., ... & Keutzer, K. (2020, April). Q-bert: Hessian based ultra low precision quantization of bert. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 05, pp. 8815-8821).

[14] Zafrir, Ofir, et al. "Q8bert: Quantized 8bit bert." 2019 Fifth Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS Edition (EMC2-NIPS). IEEE, 2019.

[15] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[16] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research, 21(140), 1-67.

[17] Wu, Z., Wang, S., Gu, J., Khabsa, M., Sun, F., & Ma, H. (2020). Clear: Contrastive learning for sentence representation. arXiv preprint arXiv:2012.15466.

[18] He, R., Ravula, A., Kanagal, B., & Ainslie, J. (2020). Realformer: Transformer likes residual attention. arXiv preprint arXiv:2012.11747.

[19] Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. Transactions of the association for computational linguistics, 8, 64-77.

[20] Tay, Y., Tran, V. Q., Ruder, S., Gupta, J., Chung, H. W., Bahri, D., ... & Metzler, D. (2021). Charformer: Fast character transformers via gradient-based subword tokenization. arXiv preprint arXiv:2106.12672.

[21] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. arXiv preprint arXiv:1905.07129.

[22] Izsak, P., Berchansky, M., & Levy, O. (2021). How to train BERT with an academic budget. arXiv preprint arXiv:2104.07705.

[23] Li, Xianming, and Jing Li. "Angle-optimized text embeddings." arXiv preprint arXiv:2309.12871 (2023).

[24] Wu, Hongqiu, et al. "Adversarial self-attention for language understanding." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 37. No. 11. 2023.

[25] Jiang, T., Huang, S., Luan, Z., Wang, D., & Zhuang, F. (2023). Scaling sentence embeddings with large language models. arXiv preprint arXiv:2307.16645.

[26] Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., ... & Ahmed, A. (2020). Big bird: Transformers for longer sequences. Advances in neural information processing systems, 33, 17283-17297.

[27] Gao, Tianyu, Xingcheng Yao, and Danqi Chen. "Simcse: Simple contrastive learning of sentence embeddings." arXiv preprint arXiv:2104.08821 (2021).

[28] Liu, F., Jiao, Y., Massiah, J., Yilmaz, E., & Havrylov, S. (2021). Trans-encoder: Unsupervised sentence-pair modelling through self-and mutual-distillations. arXiv preprint arXiv:2109.13059.

[29] Liu, X., Yu, H. F., Dhillon, I., & Hsieh, C. J. (2020, November). Learning to encode position for transformer with continuous dynamical model. In International conference on machine learning (pp. 6327-6335). PMLR.

[30] Gao, J., Xu, H., Shi, H., Ren, X., Yu, P. L., Liang, X., ... & Li, Z. (2022, June). Autobert-zero: Evolving bert backbone from scratch. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 10, pp. 10663-10671).

[31] Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Zhou, D. (2020). Mobilebert: a compact task-agnostic bert for resource-limited devices. arXiv preprint arXiv:2004.02984.

[32] Lee-Thorp, J., Ainslie, J., Eckstein, I., & Ontanon, S. (2021). Fnet: Mixing tokens with fourier transforms. arXiv preprint arXiv:2105.03824.

[33] Iandola, F. N., Shaw, A. E., Krishna, R., & Keutzer, K. W. (2020). SqueezeBERT: What can computer vision teach NLP about efficient neural networks?. arXiv preprint arXiv:2006.11316.

[34] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., ... & Liu, Q. (2019). Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351.

[35] Tsuji, K., Hiraoka, T., Cheng, Y., & Iwakura, T. (2024). SubRegWeigh: Effective and Efficient Annotation Weighing with Subword Regularization. arXiv preprint arXiv:2409.06216.

[36] Liu, X., He, P., Chen, W., & Gao, J. (2019). Improving multi-task deep neural networks via knowledge distillation for natural language understanding. arXiv preprint arXiv:1904.09482.

[37] Ratner, A., Hancock, B., Dunnmon, J., Sala, F., Pandey, S., & RÃ©, C. (2019, July). Training complex models with multi-task weak supervision. In Proceedings of the AAAI conference on artificial intelligence (Vol. 33, No. 01, pp. 4763-4771).

[38] Ji, Y., & Eisenstein, J. (2013, October). Discriminative improvements to distributional sentence similarity. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 891-896).

[39] Fei, Hao, et al. "Conversational semantic role labeling with predicate-oriented latent graph." arXiv preprint arXiv:2210.03037 (2022).

[40] Wu, H., Xu, K., Song, L., Jin, L., Zhang, H., & Song, L. (2021). Domain-adaptive pretraining methods for dialogue understanding. arXiv preprint arXiv:2105.13665.

[41] Wu, Han, Kun Xu, and Linqi Song. "CSAGN: Conversational structure aware graph network for conversational semantic role labeling." arXiv preprint arXiv:2109.11541 (2021).

[42] Xu, K., Wu, H., Song, L., Zhang, H., Song, L., & Yu, D. (2021). Conversational semantic role labeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 2465-2475.

[43] C. Muniyappa and E. Kim, "Evolutionary Algorithms Approach For Search Based On Semantic Document Similarity," in *ICCCM '23: Proceedings of the 2023 11th International Conference on Computer and Communications Management*, arXiv:2502.15348, 2025.

[44] Y. Zhang, F. Wei, J. Li, Y. Wang, Y. Yu, J. Chen, Z. Cai, X. Liu, W. Wang, P. Wang, and Z. Wang, "Constructing a Norm for Children's Scientific Drawing: Distribution Features Based on Semantic Similarity of Large Language Models," arXiv:2502.14620, 2025.

[45] X. Pan, "Exploring RWKV for Sentence Embeddings: Layer-wise Analysis and Baseline Comparison for Semantic Similarity," arXiv:2502.05704, 2025.

[46] K. Zhou, H. Gao, S. Chen, D. Edelstein, D. Jurafsky, and C. Shani, "Rethinking Word Similarity: Semantic Similarity through Classification Confusion," in *NAACL-main-2025*, arXiv:2502.03721, 2025.

[47] L. Gao, Z. Liu, and Q. Zhang, "A Comprehensive Framework for Semantic Similarity Analysis of Human and AI-Generated Text Using Transformer Architectures and Ensemble Techniques," arXiv:2501.09538, 2025.

[48] Q. Pan, W. Qiao, J. Lou, B. Ji, and S. Li, "DuSSS: Dual Semantic Similarity-Supervised Vision-Language Model for Semi-Supervised Medical Image Segmentation," arXiv:2411.13616, 2024.

[49] V. Vančura, P. Kordík, and M. Straka, "beeFormer: Bridging the Gap Between Semantic and Interaction Similarity in Recommender Systems," in *RecSys 2024*, arXiv:2408.06643, 2024.

[50] X. Li, J. Lipp, A. Shakir, R. Huang, and J. Li, "BM$\chi$: Entropy-weighted Similarity and Semantic-enhanced Lexical Search," arXiv:2407.21139, 2024.

[51] O. Nacar and A. Koubaa, "Enhancing Semantic Similarity Understanding in Arabic NLP with Nested Embedding Learning," arXiv:2407.19173, 2024.

[52] S. M. Sadjadi, Z. Rajabi, L. Rabiei, and M.-S. Moin, "FarSSiBERT: A Novel Transformer-based Model for Semantic Similarity Measurement of Persian Social Networks Informal Texts," arXiv:2407.12950, 2024.

[53] D. Qashqai, E. Mousavian, S. B. Shokouhi, and S. Mirzakuchaki, "CSFNet: A Cosine Similarity Fusion Network for Real-Time RGB-X Semantic Segmentation of Driving Scenes," arXiv:2406.19413, 2024.

[54] H. Waghela, J. Sen, and S. Rakshit, "Saliency Attention and Semantic Similarity-Driven Adversarial Perturbation," in *5th International Conference on Data Science and Applications, Jaipur, India*, arXiv:2406.11441, 2024.

[55] B. Zhang and C. Li, "Pcc-tuning: Breaking the Contrastive Learning Ceiling in Semantic Textual Similarity," in *EMNLP 2024*, arXiv:2406.05326, 2024.

[56] B. Zhang and C. Li, "Advancing Semantic Textual Similarity Modeling: A Regression Framework with Translated ReLU and Smooth K2 Loss," in *EMNLP 2024*, arXiv:2406.04438, 2024.

[57] W. Ansar, S. Goswami, and A. Chakrabarti, "TexIm FAST: Text-to-Image Representation for Semantic Similarity Evaluation using Transformers," arXiv:2406.03865, 2024.

[58] S. Fan, Z. Bao, C. Dong, H. Liang, X. Xu, and P. Zhang, "Semantic Similarity Score for Measuring Visual Similarity at Semantic Level," arXiv:2406.03673, 2024.

[59] J. Tu, K. Xu, L. Yue, B. Ye, K. Rim, and J. Pustejovsky, "Linguistically Conditioned Semantic Textual Similarity," in *ACL 2024*, arXiv:2406.03158, 2024.

[60] S. Ao, S. Rueger, and A. Siddharthan, "CSS: Contrastive Semantic Similarity for Uncertainty Quantification of LLMs," in *Conference on Uncertainty in Artificial Intelligence (UAI) 2024*, arXiv:2405.20003, 2024.

[61] A. Nikitin, J. Kossen, Y. Gal, and P. Marttinen, "Kernel Language Entropy: Fine-grained Uncertainty Quantification for LLMs from Semantic Similarities," arXiv:2405.19831, 2024.

[62] Z. Zhang, N. Wang, H. Li, and Z. Wang, "Similarity Guided Multimodal Fusion Transformer for Semantic Location Prediction in Social Media," arXiv:2405.05143, 2024.

[63] E. Hosonuma, T. Yamazaki, T. Miyoshi, A. Taya, Y. Nishiyama, and K. Sezaki, "Image Generative Semantic Communication with Multi-Modal Similarity Estimation for Resource-Limited Networks," in *IEICE Transactions on Communications*, arXiv:2404.06124, 2024.

[64] Z. Kachwala, J. An, H. Kwak, and F. Menczer, "Rematch: Robust and Efficient Matching of Local Knowledge Graphs to Improve Structural and Semantic Similarity," in *NAACL 2024 Proceedings*, arXiv:2404.01740, 2024.

[65] T. Mahmud, S. Amizadeh, K. Koishida, and D. Marculescu, "Weakly-supervised Audio Separation via Bi-modal Semantic Similarity," in *ICLR 2024*, arXiv:2403.14441, 2024.

[66] L. Köberlein, D. Probst, and R. Lenz, "Quantifying Semantic Query Similarity for Automated Linear SQL Grading: A Graph-based Approach," in *BTW 2025*, arXiv:2403.14341, 2024.

[67] J. Liu, Y. Yang, and K. Y. Tam, "Beyond Surface Similarity: Detecting Subtle Semantic Shifts in Financial Narratives," arXiv:2403.08799, 2024.

[68] S. Xiao, Y. Chen, Y. Song, L. Chen, L. Sun, Y. Zhen, and Y. Chang, "UI Semantic Group Detection: Grouping UI Elements with Similar Semantics in Mobile Graphical User Interface," in *Displays*, arXiv:2403.04212, 2024.

[69] S. Han, S. J. Park, C. W. Kim, and Y. M. Ro, "Persona Extraction Through Semantic Similarity for Emotional Support Conversation Generation," in *ICASSP 2024*, arXiv:2403.00290, 2024.

[70] B. A. Madhabhavi, G. Karevvanavar, R. V. Bhat, and N. Pappas, "Semantic Text Transmission via Prediction with Small Language Models: Cost-Similarity Trade-off," arXiv:2402.14888, 2024.

[71] R. Petcu and S. Maji, "Efficient Data Selection Employing Semantic Similarity-based Graph Structures for Model Training," in *ICML 2023 Workshop: Sampling and Optimization in Discrete Space*, arXiv:2402.13130, 2024.

[72] I. Rep, D. Dukić, and J. Šnajder, "Are ELECTRA's Sentence Embeddings Beyond Repair? The Case of Semantic Textual Similarity," in *EMNLP 2024 Findings*, arXiv:2402.11908, 2024.

[73] S. G. Picha, D. A. Chanti, and A. Caplier, "Semantic Textual Similarity Assessment in Chest X-ray Reports Using a Domain-Specific Cosine-Based Metric," arXiv:2402.11398, 2024.

[74] S. Xu, Z. Wu, H. Zhao, P. Shu, Z. Liu, W. Liao, S. Li, A. Sikora, T. Liu, and X. Li, "Reasoning before Comparison: LLM-Enhanced Semantic Similarity Metrics for Domain Specialized Text Analysis," arXiv:2402.10567, 2024.

[75] L. Yu, B. Liu, Q. Lin, X. Zhao, and C. Che, "Semantic Similarity Matching for Patent Documents Using Ensemble BERT-related Model and Novel Text Processing Method," in *MLMI 2023*, arXiv:2401.04422, 2024.

[76] T. vor der Brück and M. Pouly, "Estimating Text Similarity based on Semantic Concept Embeddings," in *IARIA Congress Proceedings, 2023*, arXiv:2312.09006, 2024.

[77] L. Yi, H. Yu, Z. Shi, G. Wang, X. Liu, L. Cui, and X. Li, "FedSSA: Semantic Similarity-based Aggregation for Efficient Model-Heterogeneous Personalized Federated Learning," in *IJCAI 2024*, arXiv:2311.04927, 2023.

[78] J. Gatto, O. Sharif, P. Seegmiller, P. Bohlman, and S. M. Preum, "Text Encoders Lack Knowledge: Leveraging Generative LLMs for Domain-Specific Semantic Textual Similarity," in *GEM@EMNLP-2023*, arXiv:2309.04607, 2023.

[79] E. Kennedy, S. Vadlamani, H. M. Lindsey, K. S. Peterson, K. D. O'Connor, K. Murray, R. Agarwal, H. H. Amiri, R. K. Andersen, T. Babikian, et al., "Linking Symptom Inventories using Semantic Textual Similarity," arXiv:2309.03911, 2023.

[80] X. Xu, Z. Xuan, S. Feng, S. Cheng, Y. Ye, Q. Shi, G. Tao, L. Yu, Z. Zhang, and X. Zhang, "PEM: Representing Binary Program Semantics for Similarity Analysis via a Probabilistic Execution Model," arXiv:2308.12842, 2023.

[81] A. Venkataramanan, M. Laviale, and C. Pradalier, "Integrating Visual and Semantic Similarity Using Hierarchies for Image Retrieval," in *ICVS 2023*, arXiv:2308.07429, 2023.

[82] C.-Y. Su and C. McMillan, "Semantic Similarity Loss for Neural Source Code Summarization," in *Journal of Software Evolution and Process*, arXiv:2308.07359, 2023.

[83] J. J. Schneider, M. Adler, C. Ammer-Herrmenau, A. O. König, U. Sax, and J. Hügel, "Improving ICD-based semantic similarity by accounting for varying degrees of comorbidity," arXiv:2308.05896, 2023.

[84] Z. Liu, Z. Zhang, S. Ma, D. Liu, J. Zhang, C. Chen, S. Liu, M. E. Ahmed, and Y. Xiang, "SemDiff: Binary Similarity Detection by Diffing Key-Semantics Graphs," arXiv:2308.00157, 2023.

[85] F. Remy, S. Scaboro, and B. Portelli, "Boosting Adverse Drug Event Normalization on Social Media: General-Purpose Model Initialization and Biomedical Semantic Text Similarity Benefit Zero-Shot Linking in Informal Contexts," arXiv:2307.16638, 2023.

[86] M. Bocharova, E. Malakhov, and V. Mezhuyev, "VacancySBERT: The Approach for Representation of Titles and Skills for Semantic Similarity Search in the Recruitment Domain," in *Applied Aspects of Information Technology, 1(6), 52-59 (2023)*, arXiv:2307.10639, 2023.

[87] N. L. Le, M.-H. Abel, and P. Gouspillou, "Improving Semantic Similarity Measure Within a Recommender System Based-on RDF Graphs," in *ICITS 2023*, arXiv:2307.09758, 2023.

[88] A. Nicolson, J. Dowling, and B. Koopman, "Longitudinal Data and a Semantic Similarity Reward for Chest X-Ray Report Generation," arXiv:2307.09274, 2023.

[89] J. Zang and H. Liu, "Improving Text Semantic Similarity Modeling through a 3D Siamese Network," arXiv:2307.07851, 2023.

[90] T. Schopf, E. Gerber, M. Ostendorff, and F. Matthes, "AspectCSE: Sentence Embeddings for Aspect-based Semantic Textual Similarity Using Contrastive Learning and Structured Knowledge," in *RANLP 2023*, arXiv:2307.00925, 2023.

[91] J. Martinez-Gil, "Automatic Design of Semantic Similarity Ensembles Using Grammatical Evolution," arXiv:2306.17810, 2023.

[92] Mohebbi, Majid, Seyed Naser Razavi, and Mohammad Ali Balafar. "Computing semantic similarity of texts based on deep graph learning with ability to use semantic role label information." Scientific reports 12.1 (2022): 14777.