English Semantic Textual Similarity and Search De-duplication

DIAO Daokui
JIN Zhuoyuan
ZHU Yongchao

Feb 2026

Outline

Introduction: STS and Search De-duplication

- **Importance in Systems**: Semantic Textual Similarity (STS) and de-duplication are critical for retrieval and question-answering systems.
- **Information Redundancy**: Pervasive on platforms like Quora, where users express the same intent using diverse phrasing.
- **Core System Task**: Systems must judge whether two texts are equivalent in intent/logic and determine the degree of similarity.
- **Accuracy Requirement**: Effective judging is essential for precise de-duplication to prevent fragmented results and redundant content.

Motivation: Fine-Grained Semantic Sensitivity

- **Current Limitation**: SBERT(Pre-trained) models are often insufficiently sensitive to fine-grained semantic conflicts.
- **The "Logic Gap"**:
  - S_1: "The meeting is cancelled."
  - S_2: "The meeting is **not** cancelled."
- **Problem**: These sentences have opposite logic, yet cosine similarity in vector space can exceed 0.9.
- **Goal**: Construct a robust model that distinguishes logical equivalence beyond broad semantic alignment.

```
Please enter two sentences (enter q to quit):
Sentence 1: The meeting is not cancelled.
Sentence 2: The meeting is cancelled.

BERT QQP→STS-B Similarity: 0.676 → Not Duplicate
SBERT STS-B Cosine Similarity: 0.926 → Duplicate
```

Methodology: Baseline Models

- **Lexical Baseline (TF-IDF)**:
  - Based on word frequency statistics; represents the performance floor.
  - Brittle under paraphrasing and lexical mismatch.
- **Static Embedding (Word2Vec)**:
  - Uses dense vectors with Mean Pooling.
  - Tests the limits of non-contextual embeddings in capturing semantics.
- **Pre-trained Transformers**:
  - **Bi-Encoders (SBERT)**: Efficient independent embeddings, but lacks word-level interaction for logic.
  - **Cross-Encoders (BERT/ROBERTa)**: Deep self-attention allows full interaction across sentence pairs.

Methodology: Detailed STL-BERT Pipeline

Architecture: BERT Cross-Encoder

- **Input**: $[CLS] + S_1 + [SEP] + S_2 + [SEP]$
- **Interaction**: Deep self-attention performs full word-level interaction between sentences at every layer.
- **Advantage**: Find slight differences (e.g. cancelled/not cancelled) that independent pooling (Bi-Encoders) ignores.

Sequential Transfer Learning

- **Phase A (Logic First)**: Pre-training on QQP to teach model a clear logical decision boundary.
- **Phase B (Perception Later)**: Fine-tuning on STS-B to produce a continuous similarity score.

Hyperparameter	Phase A: QQP (Classification)	Phase B: STS-B (Regression)
Learning Rate	2×10^{-5}	1×10^{-5}
Batch Size (Train/Eval)	16 / 32	16 / 32
Epochs	3	3
Max Sequence Length	128	128
Weight Decay	0.01	0.01
Optimizer	AdamW	AdamW
Head Type	2-Class Classification	1-D Regression

Overall Performance Comparison

Model	QQP Acc	QQP F1	STS-B Pearson	STS-B Spearman
TF-IDF	0.684	0.347	0.565	0.558
Word2Vec	0.642	0.616	0.635	0.628
BERT (Pre-trained)	0.376	0.534	-	-
SBERT (Pre-trained)	0.736	0.713	0.866	0.839
BERT (STS-B)	0.614	0.634	0.844	0.804
SBERT (STS-B)	0.750	0.721	0.889	0.866
STL-BERT (QQP→STS-B)	**0.922**	**0.891**	**0.900**	**0.875**

Table: Overall performance comparison.

STL-BERT (QQP→STS-B) is best on both QQP and STS-B.

Robustness Analysis: Custom Stress-Test Set

- **Fine-Grained Stress-Test Set**:
  - 32 manually crafted English sentence pairs.
  - 8 perturbation types: Negation, Role Swap, Numeric, Modal Shift, etc.
- **Objective**: Explicitly examine whether models distinguish lexically similar pairs that differ in crucial semantic aspects.

	cat	s1	s2	label
0	negation	The product is available.	The product is available.	1
1	negation	The product is available.	The product is not available.	0
2	negation	He likes coffee.	He does not like coffee.	0
3	negation	The service did not fail.	The service was successful.	1
4	inc_dec	Revenue increased by 10 percent.	Revenue went up by 10 percent.	1
5	inc_dec	Revenue increased by 10 percent.	Revenue decreased by 10 percent.	0
6	inc_dec	The temperature rose rapidly.	The temperature fell rapidly.	0
7	inc_dec	Sales dropped this quarter.	Sales decreased this quarter.	1
8	comparative	This model is more accurate.	This model is less accurate.	0
9	comparative	Version A is better than version B.	Version A outperforms version B.	1
10	comparative	The new phone is cheaper.	The new phone is more expensive.	0
11	comparative	Latency is lower in system X.	System X has lower latency.	1
12	role_swap	Alice gave Bob the book.	Bob gave Alice the book.	0

Robustness Analysis: Results

Key Findings:

- STL-BERT (QQP→STS-B) consistently outperforms SBERT under fine-grained stress conditions.
- Sequential transfer improves sensitivity to modal, directional, and comparative shifts.
- Negation and role swap remain challenging for both models.

Model	Acc	F1	Pearson	Spearman	MAE
BERT (QQP→STS-B)	**0.650**	**0.632**	**0.342**	**0.322**	**0.507**
SBERT-STS-B	0.550	0.571	0.200	0.235	0.528

Table: Aggregated robustness results

Category	BERT (QQP→STS-B)	SBERT-STS-B
Quantifier	1.00	1.00
Modal	1.00	0.50
Direction	1.00	0.67
Comparative	0.67	0.33
Numeric	0.67	0.67
Negation	0.50	0.50
Role Swap	0.33	0.33
Inc/Dec	0.33	0.67

Table: Accuracy by perturbation type

Reference

- Devlin et al., 2019. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.
- Reimers & Gurevych, 2019. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*.
- Cer et al., 2017. *SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation*.
- Zhang et al., 2019. *PAWS: Paraphrase Adversaries from Word Scrambling*.
- Phang et al., 2018. *STILTs: Sentence-level Training on Intermediate Labeled-data Tasks*.
- Wang et al., 2018. *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*.

Contributions

- **JIN Zhuoyuan:** Design and implementation of the STL-BERT model, design and construction of the custom dataset, overall performance comparison, robustness analysis, and writing the Methodology, Simulation, and Result sections.
- **DIAO Daokui:** Training and fine-tuning of STL-BERT, development of baselines and single-task fine-tuned models, performance testing and comparisons, writing the Abstract, Introduction, and Data sections, creating experimental charts, and organizing references.
- **ZHU Yongchao:** Implementation of baseline models, testing different models on the test sets, related work research, summarizing fine-grained failure cases of traditional STS methods and Transformer models, and writing the Related Work and Motivation sections.