Indian Institute of Technology, Kharagpur

SciSearch

Query-by-Example for Scientific Article Retrieval

Ashutosh Kumar Singh | Suhas Jain | Ashwamegh Rathore | Nakul Aggarwal

CS60092 - Information Retrieval

Overview

Problem Statement

Retrieval of facet wise most similar research papers.

User Input:
Query Paper: Paper title and abstract
Facet: Background / Method / Result

Results:
Ranked list of most similar papers from a research paper corpus.

Dataset Used:
CSFCube Dataset (2021)

Paper: https://arxiv.org/pdf/2103.12906.pdf

GitHub: https://github.com/iesl/CSFCube

Objectives

Making Literature Reviews Easy

Literature Reviews often require spending hours on the internet finding papers which have a similar background/method/result. We want to make that process effortless.

Checking Novelty

SciSearch can be used to see if there exist papers which have used a particular method before. As the amount of research grows with time, this task becomes more difficult.

Comparing Results

A vital part of research is comparing results with other papers which worked on same/similar problem statement. SciSearch can make finding such competing papers easier.

Motivation

Better Semantic Similarity

Most of the works till now use cosine similarity or L2 distance based measures, which fail to capture semantic relatedness. We aim to capture the semantic similarity with SciSearch.

Better Mechanistic Similarity

Method is the most challenging facet of the three as it often relies on determining similarity across a sequence of actions. With SciSearch we aim to improve in this domain.

Domain Specific Similarity

It is very important to treat concepts like “stacking”, “ensemble strategy”, and “bagging” as related. Existing search engines lack in this domain.

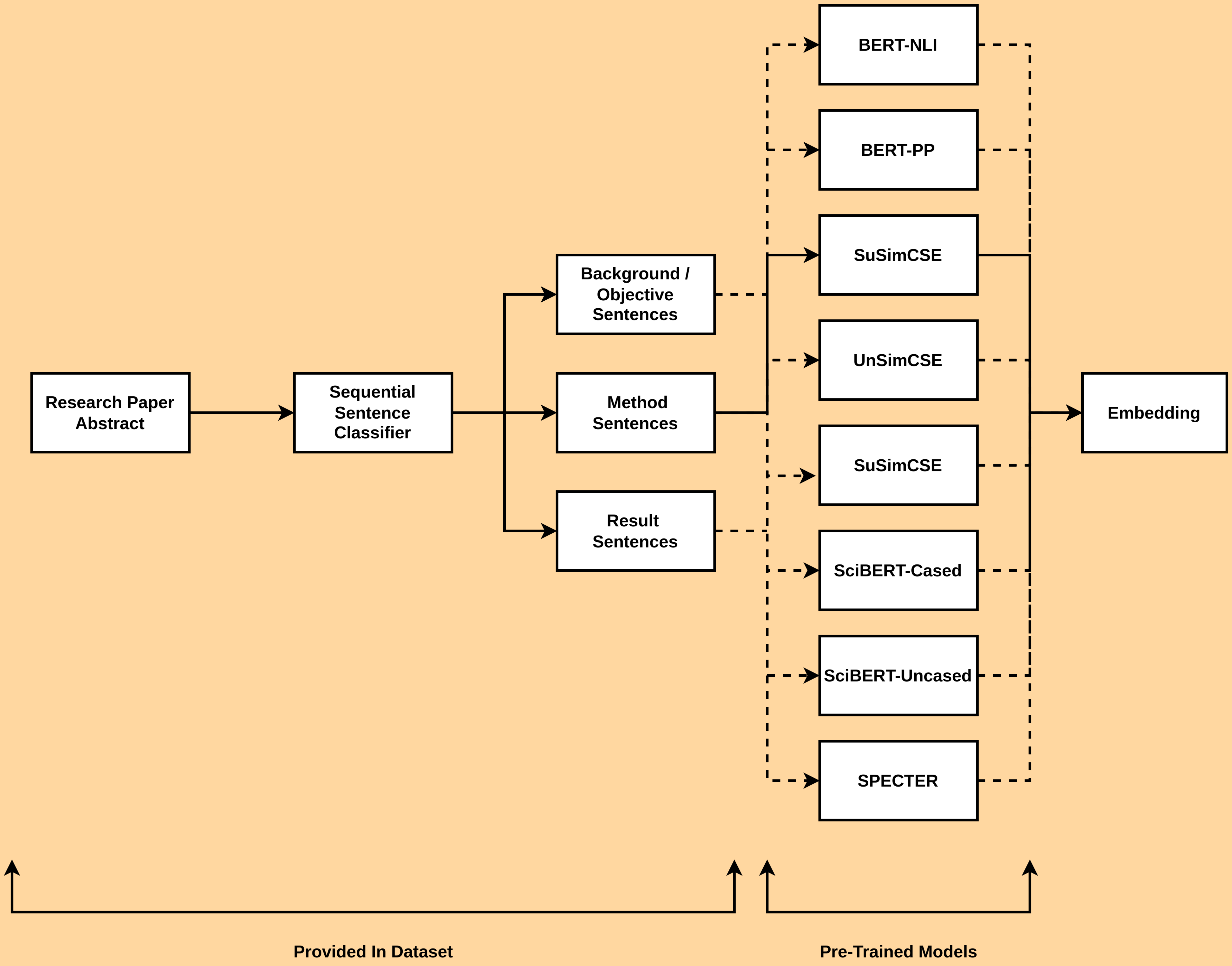
CSFCube Dataset

800,000 computer science papers, sourced out of total corpus of 81.1M papers from S2ORC. Used for training, testing, validation and demonstration in the project

- A human annotated relevance score between 0-3 provided for each query-document pair.
- Each query has a candidate pool of 100-200 research papers.
- Currently contains only abstract and title of each paper.
- Contains 16 background queries, 17 method queries, and 17 result queries.
- A total of 6244 query-candidate pairs.

Techniques and Experiments

Feature Extraction



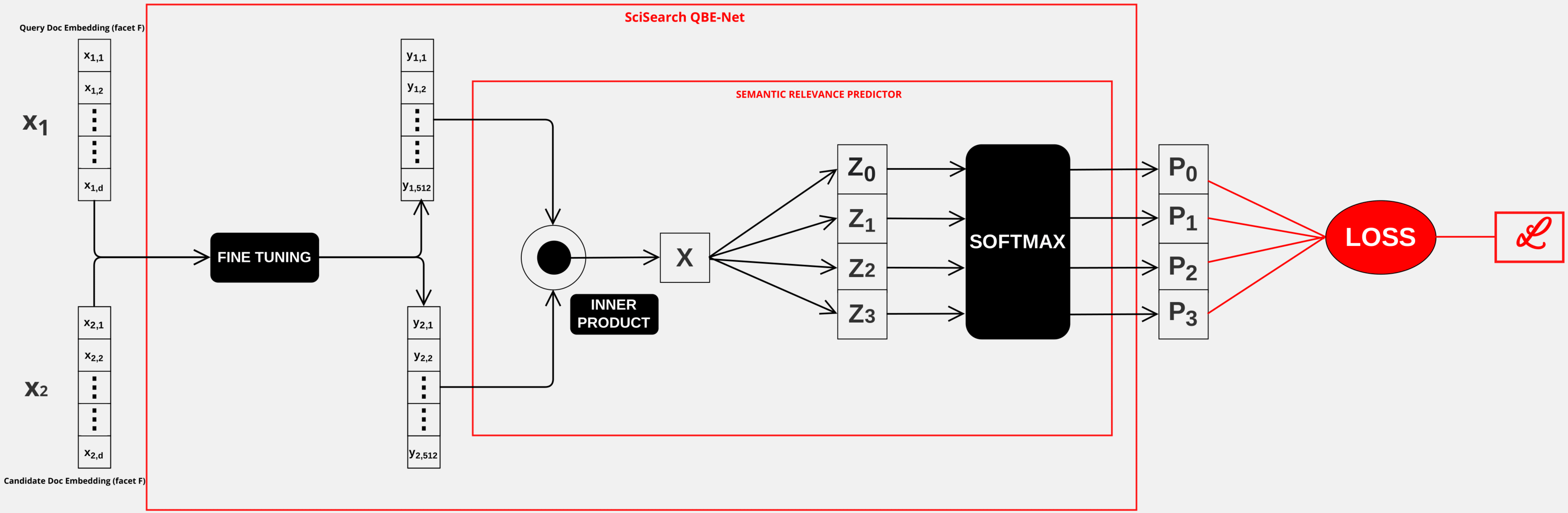
Neural Network Architecture

Fine Tuning Network

- To tweak the knowledge gained by these pre-trained models
- Responsible for this transfer of learning from one domain to the other
- Simple fine-tuning neural network:
  - Fully-connected linear layers with linear bias
  - TanH activation

Semantic Relevance Predictor

- Learns the relevance score (0-3) between any two representations
- Maps the dot-product to a 4-dimensional vector of logits by passing it through a linear layer
- Softmax applied to get probabilities of labels 0-3
- Better than using a non-flexible metric like dot-product to gauge the relevance

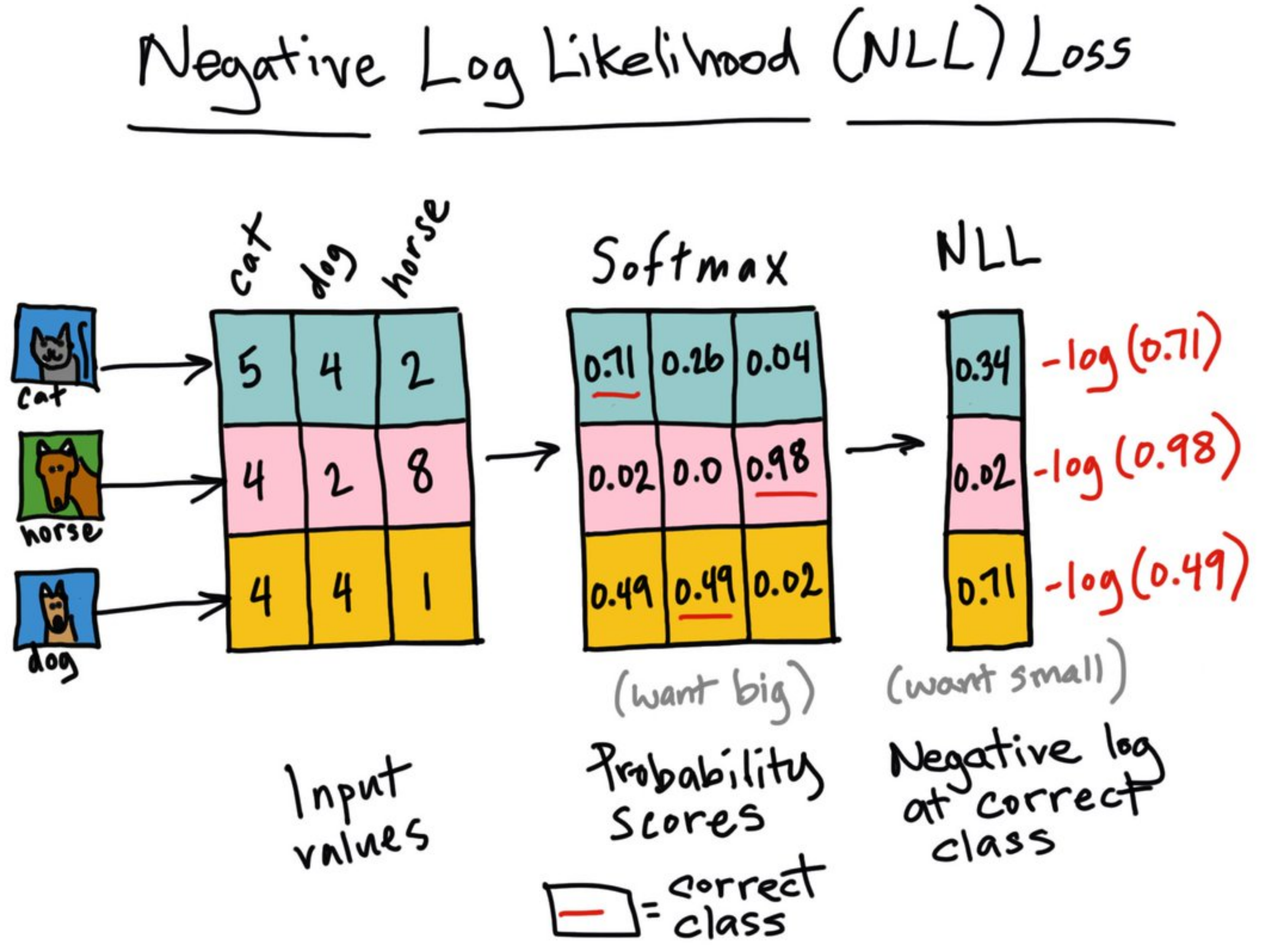


SciSearch QBE-Net
Query Doc Embedding (facet F)
x_1
$x_{1,1}$
$x_{1,2}$
$x_{1,d}$
x_2
$x_{2,1}$
$x_{2,2}$
$x_{2,d}$
Candidate Doc Embedding (facet F)
FINE TUNING
$y_{1,1}$
$y_{1,2}$
$y_{1,512}$
$y_{2,1}$
$y_{2,2}$
$y_{2,512}$
SEMANTIC RELEVANCE PREDICTOR
INNER PRODUCT
X
Z_0
Z_1
Z_2
Z_3
SOFTMAX
P_0
P_1
P_2
P_3
LOSS
$\mathscr{L}$

Selecting the Loss Function

Negative Log Likelihood Loss

Simple multi-class classification problem in hand, so naturally we select NLL-Loss.



Problems?

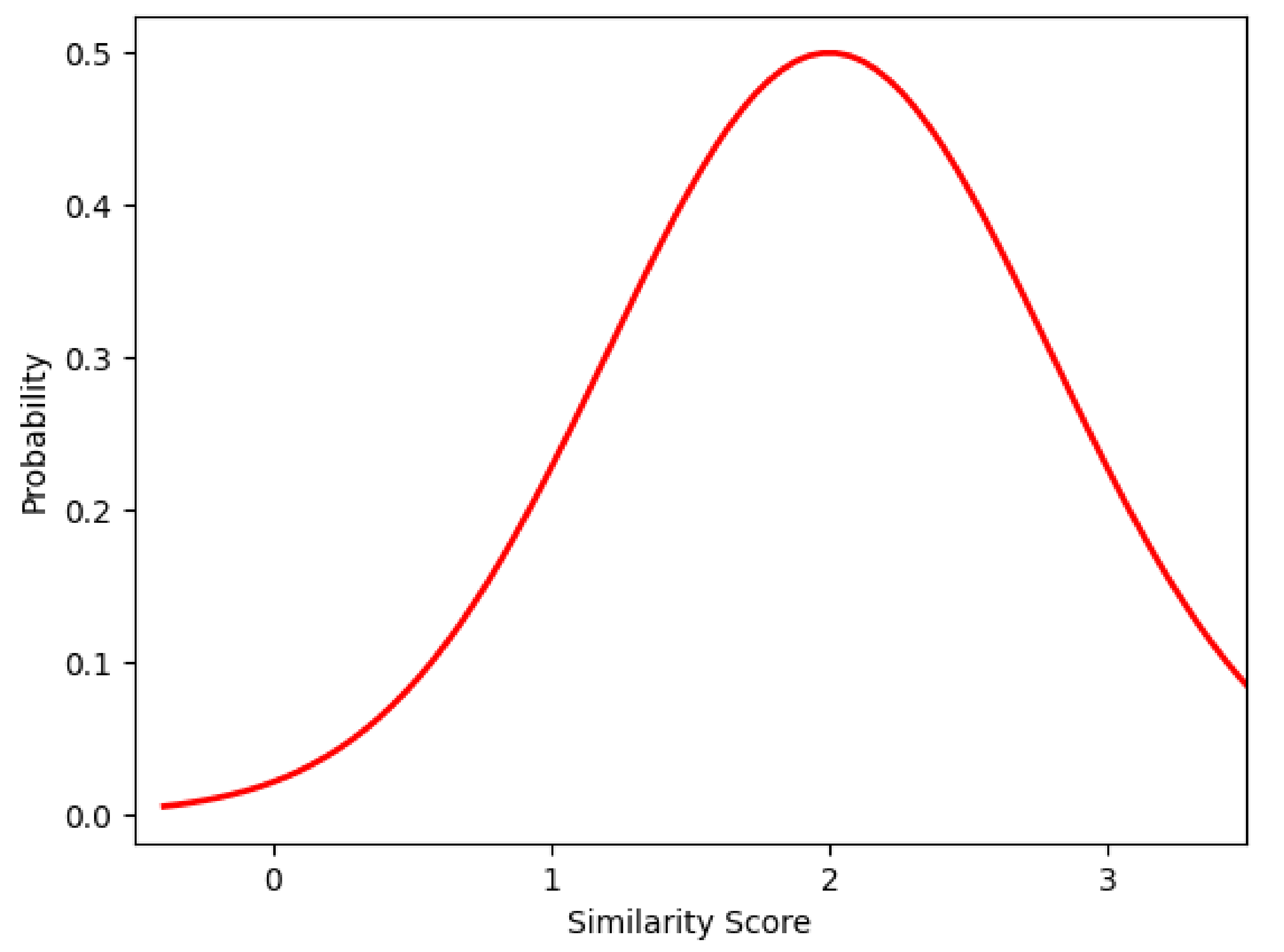
A simple thought experiment:
Assume the ground truth label is 2. Loss function only cares about probability of label 2, treats all other classes as same.

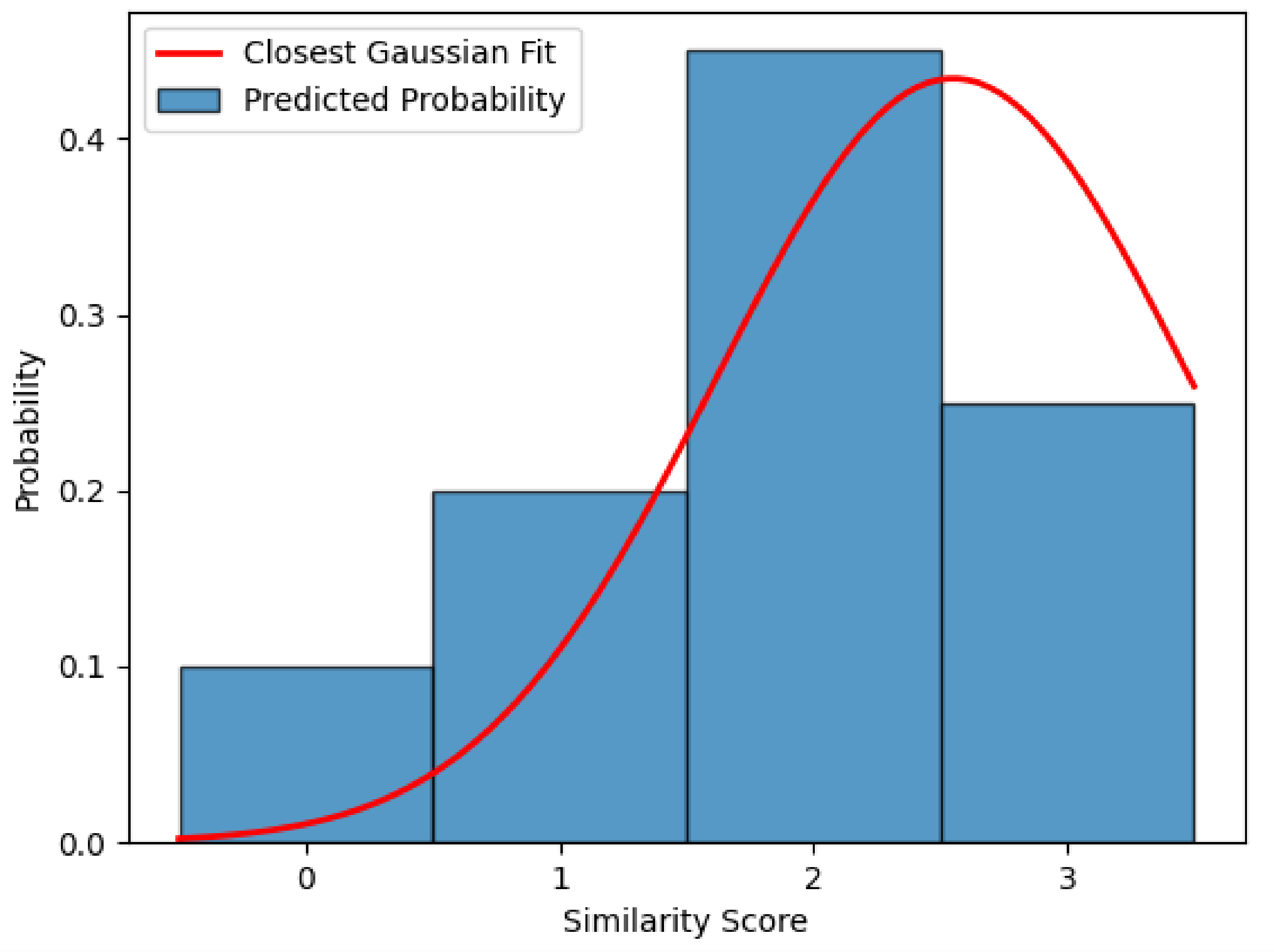
Ask question:
Is model predicting label as 0 and model predicting label as 1 or 3 same?

Kullback-Leibler Divergence Loss

Measures difference between two probability distributions. The ground truth and predicted results are converted to probability distributions as following:

- **Ground Truth:** Gaussian distribution centred around the ground truth label.
- **Predicted Value:** A best fit Gaussian distribution from probabilities obtained from softmax layer.



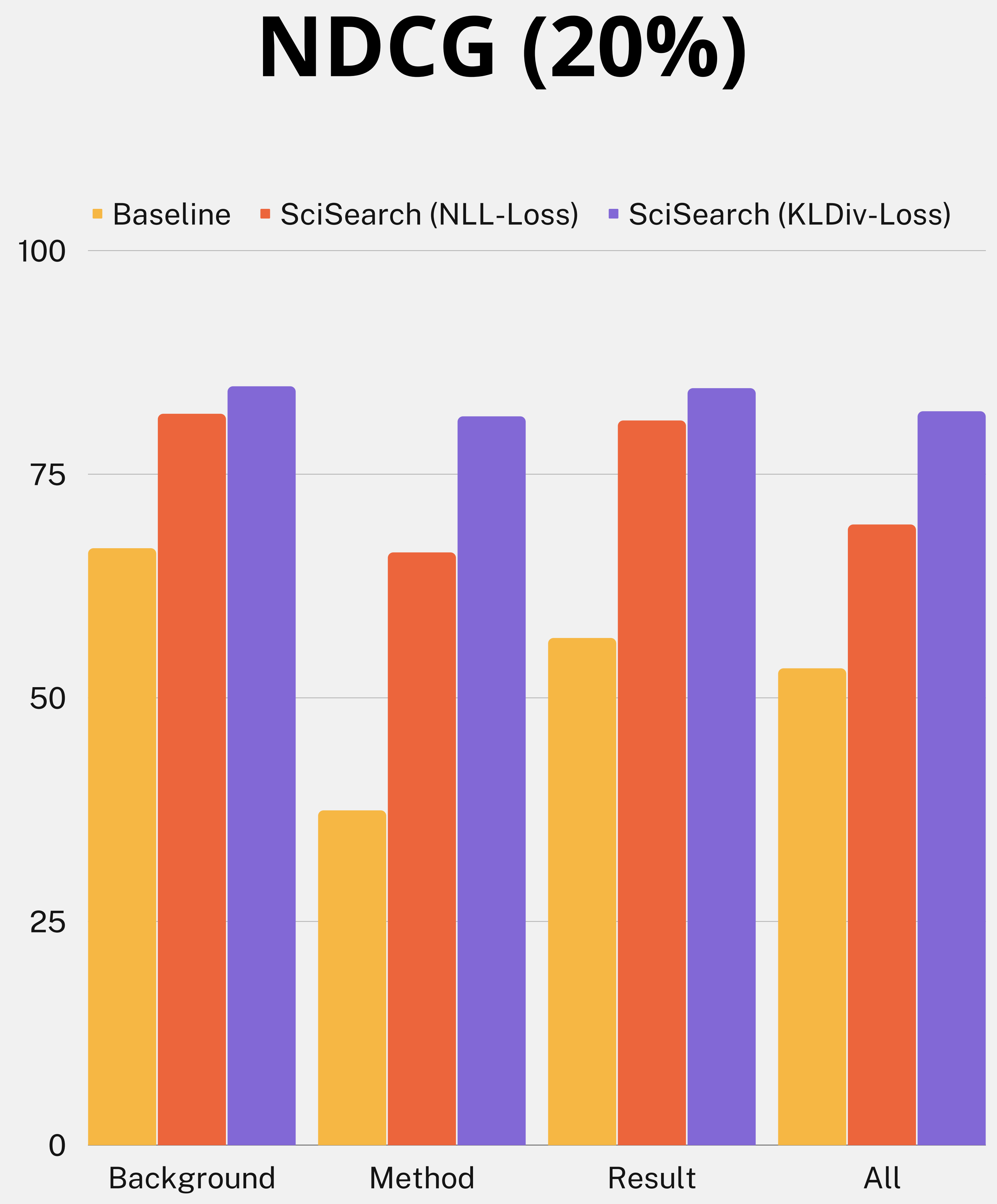


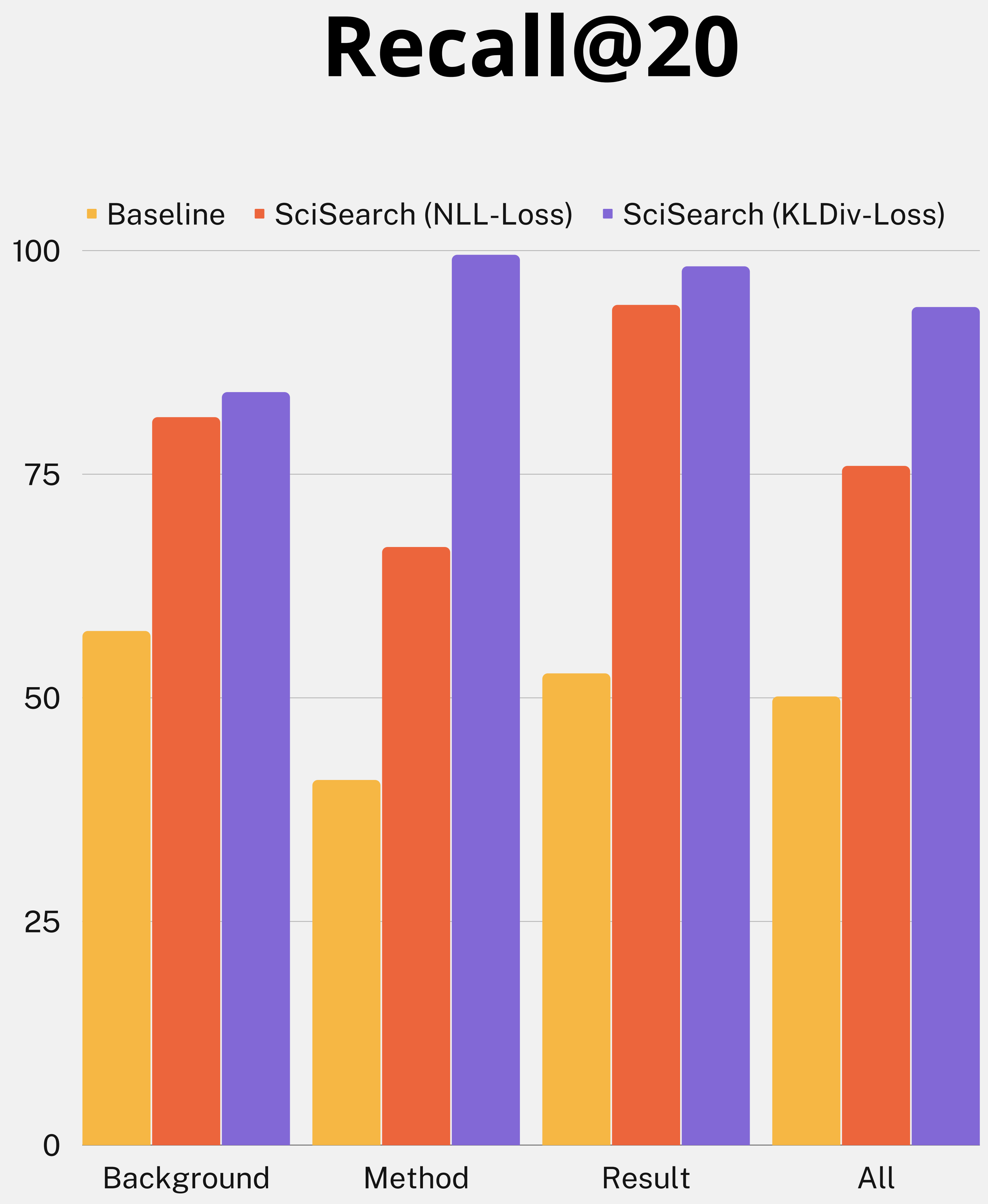
Results

We get significant growth in all evaluation parameters across facets.

Key observations are:

- Both loss functions provide significant improvement over the baseline results
- Biggest jump is seen in case of method facet, one of our primary objectives
- Among different parameters highest jump is seen in recall, implying we are retrieving most of the relevant documents
- NDCG at both 20% and 100% is much higher when compared to baseline. This indicates we produce superior ranking amongst the retrieved documents


NDCG (20%)
Baseline
SciSearch (NLL-Loss)
SciSearch (KLDiv-Loss)
100
75
50
25
0
Background
Method
Result
All


Recall@20
Baseline
SciSearch (NLL-Loss)
SciSearch (KLDiv-Loss)
100
75
50
25
0
Background
Method
Result
All

Demonstration

Future Work

Clean and Augment Dataset

The data contains special unicode characters and latex snippets. These can be removed or modified into natural language.

Expanding to Full S2ORC Corpus and Further

Pre-calculate embeddings and efficiently make candidate pool to make a general purpose research paper query tool.

Match Full Body Text

Currently training and querying using only the title and abstract of a paper. Devise a computationally efficient strategy to use full body text.

Thank you!