


BERT MLM - testing example utterances

Guide to Tokenization and Padding with BERT: Transforming Text into Machine-Readable Data

Tokenizers are the unsung heroes of modern natural language processing (NLP). They bridge the gap between human-readable language and...

 <https://medium.com/@piyushkashyap045/guide-to-tokenization-and-padding-with-bert-transforming-text-into-machine-readable-data-5a24bf59d36b>

```
C:\Users\nat\PycharmProjects\PythonProject\.venv\Scripts\python.exe C:\Users\nat\
```

```
# Testing missing nouns that may skipped due to hesitation/uncertainty or may be n
---- Sentence 1
```

Original: He kicked a soccer [MASK] and it broke the window.

Tokenized: ['he', 'kicked', 'a', 'soccer', '[MASK]', 'and', 'it', 'broke', 'the', 'window', '.']

Token IDs: [2002, 6476, 1037, 4715, 103, 1998, 2009, 3631, 1996, 3332, 1012]

Top predictions to replace masked token:

Token: ball, Score: 0.9098

Token: bat, Score: 0.0215

Token: stick, Score: 0.0213

Token: cap, Score: 0.0050

Token: football, Score: 0.0046

Unmasked sentence with top prediction:

He kicked a soccer ball and it broke the window.

```
# Aphasia patients struggle with "small words" such as pronouns and articles so a st
---- Sentence 2
```

Original: It is raining so he brought [MASK] umbrella.

Tokenized: ['it', 'is', 'raining', 'so', 'he', 'brought', '[MASK]', 'umbrella', '.']

Token IDs: [2009, 2003, 24057, 2061, 2002, 2716, 103, 12977, 1012]

Top predictions to replace masked token:

Token: an, Score: 0.5922

Token: his, Score: 0.2837

Token: the, Score: 0.1050

Token: my, Score: 0.0056

Token: her, Score: 0.0034

Unmasked sentence with top prediction:

It is raining so he brought an umbrella.

Generating a noun without any wider context.

---- Sentence 3

Original: She told him to bring his [MASK] as it is raining.

Tokenized: ['she', 'told', 'him', 'to', 'bring', 'his', '[MASK]', 'as', 'it', 'is', 'raining', '.']

Token IDs: [2016, 2409, 2032, 2000, 3288, 2010, 103, 2004, 2009, 2003, 24057, 101]

Top predictions to replace masked token:

Token: car, Score: 0.0901

Token: coat, Score: 0.0396

Token: horse, Score: 0.0322

Token: bicycle, Score: 0.0317

Token: bike, Score: 0.0247

Unmasked sentence with top prediction:

She told him to bring his car as it is raining.

Generating with context from a previous sentence.

---- Sentence 1

Original: It is raining so she brought her umbrella. She told him to bring his [MASK] a

Tokenized: ['it', 'is', 'raining', 'so', 'she', 'brought', 'her', 'umbrella', '.', 'she', 'told', 'him']

Token IDs: [2009, 2003, 24057, 2061, 2016, 2716, 2014, 12977, 1012, 2016, 2409, 203]

Top predictions to replace masked token:

Token: umbrella, Score: 0.8774

Token: coat, Score: 0.0102

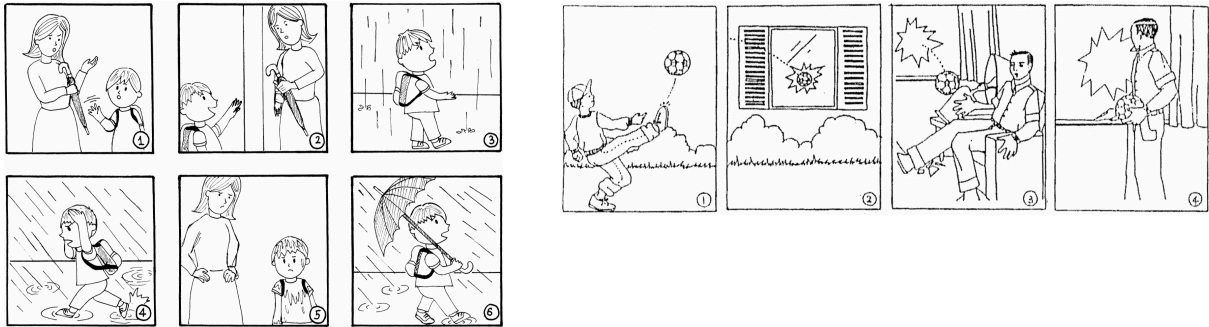
Token: car, Score: 0.0100

Token: hat, Score: 0.0077

Token: jacket, Score: 0.0065

Unmasked sentence with top prediction:

It is raining so she brought her umbrella. She told him to bring his umbrella as it is rai



— Each mask is generated simultaneously, so the second word does not take consideration of the previous chosen word.

— Ig we don't want to propagate errors so maybe keep like this

Multiple mask tokens within a sentence

Original: It is raining so she brought [MASK] umbrella. She told him to bring his [MAS

Tokenized: ['it', 'is', 'raining', 'so', 'she', 'brought', '[MASK]', 'umbrella', '!', 'she', 'told',

Token IDs: [2009, 2003, 24057, 2061, 2016, 2716, 103, 12977, 1012, 2016, 2409, 2032

Top predictions to replace masked token:

Token: her, Score: 0.4566

Token: his, Score: 0.3152

Token: an, Score: 0.1518

Token: the, Score: 0.0588

Token: my, Score: 0.0060

Top predictions to replace masked token:

Token: umbrella, Score: 0.8922

Token: coat, Score: 0.0125

Token: jacket, Score: 0.0079

Token: hat, Score: 0.0075

Token: shoes, Score: 0.0039

Unmasked sentence with top prediction:

it is raining so she brought her umbrella. she told him to bring his umbrella as it is rai

Original: He kicked a soccer [MASK] and it broke the window. The man yelled becau
Tokenized: ['he', 'kicked', 'a', 'soccer', '[MASK]', 'and', 'it', 'broke', 'the', 'window', '.',
Token IDs: [2002, 6476, 1037, 4715, 103, 1998, 2009, 3631, 1996, 3332, 1012, 1996, 2

Top predictions to replace masked token:

Token: ball, Score: 0.9094

Token: stick, Score: 0.0196

Token: bat, Score: 0.0167

Token: field, Score: 0.0067

Token: player, Score: 0.0064

Top predictions to replace masked token:

Token: window, Score: 0.9402

Token: glass, Score: 0.0443

Token: windows, Score: 0.0036

Token: windshield, Score: 0.0034

Token: door, Score: 0.0013

Unmasked sentence with top prediction:

he kicked a soccer ball and it broke the window. the man yelled because the window