

# Rough Analysis of v2\_marked\_and\_linked\_test.json

## General Information

- Num. Entries: 18
- Tokenizer used: dmis-lab/biobert-v1.1

### Definitions:

- Words represent the number of entities that have been considered distinct during the process of assigning tags to part of the text, it is essentially the text split on spaces
- Tokens are the actual pieces of data going into the model, punctuation is split from words and some large words are split down into smaller parts
- Total tags is the total number of tags across all words (If two or more words are part of the same tag this counts as a tag for each word)
- Unique tags is the total number of tags that are unique (not part of same tag)
- Max tags is the most tags a single word has
- Tagged words is the number of words that have at least one tag
- Tagged words % is the percentage of words that are tagged, it's given in brackets for individual entries.
- Avg tags is the average number of tags per word that is tagged
- MC words is the number of words with tags across multiple categories
- Total links is the total number of links across all tags (If two or more tags are linked this counts as a link for each one)
- Unique links is the total number of links that are unique (not part of the same link)
- Max links is the most amount of links a single tag or word has associated with it
- Linked tags, words is the number of tags or words that have links
- Linked % tags, words is the percentage of tags or words with links
- Avg links per tag, word is the average number of links per tag or word
- The schema used is BIES as it is all encompassing and models can convert to a lower resolution if wanted (Beginning, Inner, End, Singleton)

## Sizes

- Average words: 209.8
- Maximum words: 283
- Minimum words: 132
  
- Average tokens: 397.5
- Maximum tokens: 593
- Minimum tokens: 248
  
- Entries with over 512 tokens: 2/18, 11.11%

## Tags

### Maximums

- Total tags: 49
- Unique tags: 18

- Max tags: 1
- Tagged words: 49
- Tagged words %: 22.07%
- Avg tags: 1.0
- MC words: 0

### Averages

- Total tags: 23.83
- Unique tags: 12.39
- Max tags: 1.0
- Tagged words: 23.83
- Tagged words %: 11.56%
- Avg tags: 1.0
- MC words: 0.0

### Links

#### Maximums

- Total links: 47
- Unique links: 7
- Max links per tag, word: 6, 6
- Linked tags, words: 24, 24
- Linked % tags, words: 100.0%, 10.81%
- Avg links per tag, word: 2.76, 2.76

#### Averages

- Total links: 14.28
- Unique links: 1.94
- Max links per tag, word: 1.72, 1.72
- Linked tags, words: 10.28, 10.28
- Linked % tags, words: 48.24%, 5.07%
- Avg links per tag, word: 1.33, 1.33

### Schema

- Maximums (B, I, E, S): 14, 20, 14, 13
- Averages (B, I, E, S): 6.5, 4.94, 6.5, 5.89

### Labels

- Label: Total, Maximum per entry, Average per entry

### Categories

- Perturbing\_Action: 154, 19, 8.56
- Context: 146, 28, 8.11
- Effect: 52, 6, 2.89
- Phenotype: 77, 10, 4.28

### Perturbing\_Action

- -: 0, 0, 0.0
- Gene Loss-Of-Function: 59, 14, 3.28
- Gene Gain-Of-Function: 27, 7, 1.5
- Rnai/Knockdown: 32, 11, 1.78
- Pharmacological Inhibition: 20, 6, 1.11

- Pharmacological Augmentation: 0, 0, 0.0
- Other: 16, 7, 0.89

## Context

- -: 0, 0, 0.0
- Patient: 0, 0, 0.0
- Organism: 12, 9, 0.67
- Tissue/Organ: 6, 4, 0.33
- Neoplasm: 13, 4, 0.72
- Graft: 0, 0, 0.0
- Xenograft: 12, 4, 0.67
- Cells: 33, 8, 1.83
- Transformed Cells: 48, 14, 2.67
- Organoid: 0, 0, 0.0
- In Vitro: 6, 2, 0.33
- In Vivo: 16, 4, 0.89

## Effect

- -: 0, 0, 0.0
- Positive: 25, 4, 1.39
- Negative: 19, 4, 1.06
- Regulates: 0, 0, 0.0
- Rescues: 1, 1, 0.06
- No Effect: 7, 5, 0.39

## Phenotype

- -: 0, 0, 0.0
- Adhesion: 0, 0, 0.0
- Apoptosis: 9, 4, 0.5
- Anoikis: 0, 0, 0.0
- Autophagy: 8, 4, 0.44
- Cell Cycle Arrest: 4, 4, 0.22
- Cell Death: 0, 0, 0.0
- Cell Growth: 2, 2, 0.11
- Cell Survival: 0, 0, 0.0
- Colony Formation: 0, 0, 0.0
- Differentiation: 0, 0, 0.0
- Entosis: 0, 0, 0.0
- Epithelial-Mesenchymal Transition: 7, 4, 0.39
- Ferroptosis: 4, 4, 0.22
- Invasion: 2, 2, 0.11
- Metastasis: 11, 4, 0.61
- Migration: 3, 2, 0.17
- Mitophagy: 0, 0, 0.0
- Necroptosis: 0, 0, 0.0
- Necrosis: 0, 0, 0.0
- Oncosis: 0, 0, 0.0
- Proliferation: 8, 1, 0.44
- Pyroptosis: 0, 0, 0.0
- Quiescence: 0, 0, 0.0
- Self-Renewal: 0, 0, 0.0
- Senescence: 0, 0, 0.0
- Transformation: 0, 0, 0.0
- Tumour Growth: 13, 5, 0.72

- Tumourigenesis: 6, 2, 0.33
- Tumour Initiation: 0, 0, 0.0
- Tumour Progression: 0, 0, 0.0
- Tumour Regression: 0, 0, 0.0

## Information on each entry

**Title:** 101\_PMIID32788173.txt\_CC2.xml, Words: 177, Tokens: 346

Tag data: Total: 17, Unique: 11, Max tags: 1, Tagged words: 17 (10%), Avg tags: 1.0, MC words: 0

Link data: Total: 47, Unique: 7, Max links (tag, word): (6, 6), Linked (tags, words): (17 (100.0%), 17 (9.6%))

Avg links (tag, word): (2.76, 2.76), Schema counts (B, I, E, S): (5, 1, 5, 6)

**Title:** 586\_PMIID31342239.txt\_CC2.xml, Words: 209, Tokens: 397

Tag data: Total: 33, Unique: 18, Max tags: 1, Tagged words: 33 (16%), Avg tags: 1.0, MC words: 0

Link data: Total: 8, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (8 (24.24%), 8 (3.83%))

Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (7, 8, 7, 11)

**Title:** 491\_PMIID31101885.txt\_CC2.xml, Words: 155, Tokens: 248

Tag data: Total: 27, Unique: 11, Max tags: 1, Tagged words: 27 (17%), Avg tags: 1.0, MC words: 0

Link data: Total: 10, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (10 (37.04%), 10 (6.45%))

Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (10, 6, 10, 1)

**Title:** 155\_PMIID33372040.txt\_CC2.xml, Words: 222, Tokens: 376

Tag data: Total: 41, Unique: 18, Max tags: 1, Tagged words: 41 (18%), Avg tags: 1.0, MC words: 0

Link data: Total: 19, Unique: 2, Max links (tag, word): (2, 2), Linked (tags, words): (11 (26.83%), 11 (4.95%))

Avg links (tag, word): (1.73, 1.73), Schema counts (B, I, E, S): (13, 10, 13, 5)

**Title:** 115\_PMIID32661137.txt\_CC2.xml, Words: 222, Tokens: 375

Tag data: Total: 49, Unique: 15, Max tags: 1, Tagged words: 49 (22%), Avg tags: 1.0, MC words: 0

Link data: Total: 24, Unique: 2, Max links (tag, word): (1, 1), Linked (tags, words): (24 (48.98%), 24 (10.81%))

Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (14, 20, 14, 1)

**Title:** 522\_PMIID30680482.txt\_CC2.xml, Words: 193, Tokens: 447

Tag data: Total: 31, Unique: 16, Max tags: 1, Tagged words: 31 (16%), Avg tags: 1.0, MC words: 0

Link data: Total: 5, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (5 (16.13%), 5 (2.59%))

Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (5, 10, 5, 11)

**Title:** 140\_PMIID33239431.txt\_CC2.xml, Words: 229, Tokens: 431

Tag data: Total: 20, Unique: 11, Max tags: 1, Tagged words: 20 (9%), Avg tags: 1.0, MC words: 0

Link data: Total: 12, Unique: 2, Max links (tag, word): (2, 2), Linked (tags, words): (9 (45.0%), 9 (3.93%))

Avg links (tag, word): (1.33, 1.33), Schema counts (B, I, E, S): (8, 1, 8, 3)

**Title:** 346\_PMIID33431801.txt\_CC2.xml, Words: 269, Tokens: 593

Tag data: Total: 22, Unique: 15, Max tags: 1, Tagged words: 22 (8%), Avg tags: 1.0, MC words: 0

Link data: Total: 4, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (4 (18.18%), 4 (1.49%))

Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (2, 5, 2, 13)

**Title:** 377\_PMIID32929059.txt\_CC2.xml, Words: 270, Tokens: 533

Tag data: Total: 18, Unique: 9, Max tags: 1, Tagged words: 18 (7%), Avg tags: 1.0, MC words: 0

Link data: Total: 18, Unique: 2, Max links (tag, word): (2, 2), Linked (tags, words): (10 (55.56%), 10 (3.7%))

Avg links (tag, word): (1.8, 1.8), Schema counts (B, I, E, S): (6, 3, 6, 3)

**Title:** 481\_PMIID31570854.txt\_CC2.xml, Words: 195, Tokens: 382

Tag data: Total: 19, Unique: 10, Max tags: 1, Tagged words: 19 (10%), Avg tags: 1.0, MC words: 0

Link data: Total: 16, Unique: 2, Max links (tag, word): (2, 2), Linked (tags, words): (10 (52.63%), 10 (5.13%))

Avg links (tag, word): (1.6, 1.6), Schema counts (B, I, E, S): (5, 4, 5, 5)

**Title:** 384\_PMIID32968045.txt\_CC2.xml, Words: 226, Tokens: 366

Tag data: Total: 27, Unique: 16, Max tags: 1, Tagged words: 27 (12%), Avg tags: 1.0, MC words: 0

Link data: Total: 7, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (7 (25.93%), 7 (3.1%))

Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (9, 2, 9, 7)

**Title:** 311\_Pmid33268765.txt\_CC2.xml, Words: 223, Tokens: 424  
Tag data: Total: 22, Unique: 13, Max tags: 1, Tagged words: 22 (10%), Avg tags: 1.0, MC words: 0  
Link data: Total: 5, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (5 (22.73%), 5 (2.24%))  
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (7, 2, 7, 6)

**Title:** 486\_Pmid31802036.txt\_CC2.xml, Words: 283, Tokens: 489  
Tag data: Total: 15, Unique: 9, Max tags: 1, Tagged words: 15 (5%), Avg tags: 1.0, MC words: 0  
Link data: Total: 13, Unique: 2, Max links (tag, word): (2, 2), Linked (tags, words): (8 (53.33%), 8 (2.83%))  
Avg links (tag, word): (1.62, 1.62), Schema counts (B, I, E, S): (5, 1, 5, 4)

**Title:** 748\_Pmid32243847.txt\_CC2.xml, Words: 133, Tokens: 281  
Tag data: Total: 19, Unique: 10, Max tags: 1, Tagged words: 19 (14%), Avg tags: 1.0, MC words: 0  
Link data: Total: 14, Unique: 2, Max links (tag, word): (2, 2), Linked (tags, words): (8 (42.11%), 8 (6.02%))  
Avg links (tag, word): (1.75, 1.75), Schema counts (B, I, E, S): (4, 5, 4, 6)

**Title:** 789\_Pmid29805077.txt\_CC2.xml, Words: 132, Tokens: 280  
Tag data: Total: 6, Unique: 4, Max tags: 1, Tagged words: 6 (5%), Avg tags: 1.0, MC words: 0  
Link data: Total: 6, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (6 (100.0%), 6 (4.55%))  
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (2, 0, 2, 2)

**Title:** 15\_Pmid32249716.txt\_CC2.xml, Words: 240, Tokens: 429  
Tag data: Total: 26, Unique: 17, Max tags: 1, Tagged words: 26 (11%), Avg tags: 1.0, MC words: 0  
Link data: Total: 20, Unique: 3, Max links (tag, word): (2, 2), Linked (tags, words): (18 (69.23%), 18 (7.5%))  
Avg links (tag, word): (1.11, 1.11), Schema counts (B, I, E, S): (5, 4, 5, 12)

**Title:** 589\_Pmid31177396.txt\_CC2.xml, Words: 201, Tokens: 336  
Tag data: Total: 12, Unique: 8, Max tags: 1, Tagged words: 12 (6%), Avg tags: 1.0, MC words: 0  
Link data: Total: 7, Unique: 1, Max links (tag, word): (1, 1), Linked (tags, words): (7 (58.33%), 7 (3.48%))  
Avg links (tag, word): (1.0, 1.0), Schema counts (B, I, E, S): (3, 1, 3, 5)

**Title:** 405\_Pmid31804607.txt\_CC2.xml, Words: 198, Tokens: 422  
Tag data: Total: 25, Unique: 12, Max tags: 1, Tagged words: 25 (13%), Avg tags: 1.0, MC words: 0  
Link data: Total: 22, Unique: 3, Max links (tag, word): (2, 2), Linked (tags, words): (18 (72.0%), 18 (9.09%))  
Avg links (tag, word): (1.22, 1.22), Schema counts (B, I, E, S): (7, 6, 7, 5)