|  |  |  |
| --- | --- | --- |
| To Do | | |
| # | Task details | Assigner |
|  |  |  |
|  |  |  |
|  |  |  |

|  |  |  |
| --- | --- | --- |
| Ask stats prof. | | |
| # | Question | Assigner |
|  |  |  |
|  |  |  |
|  |  |  |

Contents

[1. Introduction 3](#_Toc160150117)

[2. Data Description 5](#_Toc160150118)

[3. Data Processing Workflow 6](#_Toc160150119)

[a. Raw Data Cleaning and Hierarchical Grouping 6](#_Toc160150120)

[b. Binary Transformation of Data Values Using Thresholds 6](#_Toc160150121)

[c. Constructing Contingency Tables 7](#_Toc160150122)

[4. Data Analysis 10](#_Toc160150123)

[a. Chi-Squared Test, Fisher's Exact Test and P-Value 10](#_Toc160150124)

[b. Odd Ration (OR) 13](#_Toc160150125)

[c. Additional Measure, Yule’s Q 16](#_Toc160150126)

[d. Accounting for Yule ’Q results in the OR results 20](#_Toc160150127)

[e. Closer Look at Contingency Tables 24](#_Toc160150128)

[f. Pointwise Mutual Information PMI 25](#_Toc160150129)

[g. Decision Making Pipeline 28](#_Toc160150130)

[5. Discussing the Findings 30](#_Toc160150131)

[Appendices 33](#_Toc160150132)

[Appendix A: Calculation Verification Sheet 33](#_Toc160150133)

[Appendix B: Overview of Code-Generated Files and Directory Structure 34](#_Toc160150134)

# Introduction

This report aims to identify the pairs of senses annotators most frequently confuse with, focusing on the senses outlined in the PDTB-3 Sense Hierarchy (figure 1) at levels 2 and 3.

We keep the level-2 senses unchanged. Furthermore, we merge the senses from level 2, which do not branch into more detailed senses at level 3, with those at level 3, considering them as part of the ‘leaves level’. Thus, the leaves level includes both the senses at level 3 and selected senses from level 2 that do not have sub-senses.

We then analyze these senses at the "leaves level" and level 2, examining each pair of senses separately. We convert the data into a binary format and apply various statistical methods to identify the pairs that demonstrate significant associations. To streamline this process, a decision pipeline was designed guide the analysis.

|  |
| --- |
| Fig 1. |
|  |

# Data Description

[[ describe the raw data that we have and how each annotator vote for one sense for each instance and that their vote is equal to 0.1]]

[[ add a note that not all the senses in the sense hierarchy table are found our raw data ]]

# Data Processing Workflow

## Raw Data Cleaning and Hierarchical Grouping

All unnecessary columns were removed from the raw data. In addition, the ‘differentcon’ and ‘norel’ columns were dropped since they were not mentioned in the Sense Hierarchy table. This leaves us with 27 sense columns for DiscoGEM and 28 sense columns for QADC. Next, senses that constitute level-2 senses were grouped together for both corpora and given the appropriate label based on the PDTB-3 Sense Hierarchy table. This resulted in four datasets that are stored under the '1\_ready\_to\_transform' directory.

## Binary Transformation of Data Values Using Thresholds

The raw data cells contain values ranging from zero to one, with a step of 0.1. These values were converted into binary, with 'V' denoting a vote and '¬V' denoting no vote so that the data can be processed similarly to word collocation in [[reference to the book]]. To this end, different thresholds were used (0.3, 0.4, and 0.5). Any value greater than or equal to the threshold is set to 'V' and '¬V' otherwise.

In other words, for α = 0.3, if 3 or more out of 10 annotators chose a certain sense for a certain instance, we consider this sense to be voted for. Otherwise (2 annotators or less), we ignored these votes.

Thus, 12 datasets were created (4 datasets \* 3 thresholds = 12 datasets). Figure 2 displays the twelve datasets to be processed. It is important to note in this report we will show the results of the two highlighted datasets; these are DiscoGEM with a threshold of 0.3 at both levels. The binary transformed datasets are stored in the ‘2\_ready\_to\_process/binary’ directory.

|  |
| --- |
| Fig 2. |
|  |

## Constructing Contingency Tables

Doing a pairwise comparison between each possible sense pair, we construct a 2x2 contingency table with four cells (Table 1).

|  |  |  |
| --- | --- | --- |
| Table 1. Contingency Table Structure | | |
|  | Sense 2 | |
| Sense 1 | V | ¬V |
| V | a | b |
| ¬V | c | d |

Where:

* a: represents the count of instances where enough annotators voted for each sense for a single instance.
* b: represents the count of instances where enough annotators voted for sense 1 but not enough annotators voted for sense 2.
* c: represents the count of instances where enough annotators voted for sense 2 but not enough annotators voted for sense 1.
* d: represents instances where not enough votes were cast for either sense 1 or sense 2.

All contingency tables can be found in the following directory: ‘3\_results/binary/contingency\_tables’.

Note that all-zero value senses were dropped from further analysis. This means that there is, at most, one cell with a zero count for each row or column in the contingency table.

Figure 3 shows a sample of the contingency table considered for the processing phase.

|  |
| --- |
| Fig 3. Sample of Accepted Contingency Tables for the DiscoGEM Dataset at the Leaves Level with α = 0.3 |
|  |

[[[ Insert how many pairs survived the all zero filtering]]].

# Data Analysis

## Chi-Squared Test, Fisher's Exact Test and P-Value

The Chi-squared value was calculated for all eligible pairs. However, in calculating the Chi-squared value, the expected table is constructed. It is recommended that all cells of the expected table have a value greater than 5 for the Chi-squared calculation to be reliable.

[[[ Insert Chi squared formula ]]]

If any cell of the expected table contains a value less than five (Figure 4), alternative methods should be considered. In this report, Fisher's exact test was used whenever the expected table had a cell with a count less than five.

|  |
| --- |
| Fig 4. Example of an Expected Table with a Value Less than Five (contrast and synchronous, DiscoGEM Dataset at the Leaves Level with α = 0.3 |
|  |

Subsequently, the P-value is calculated from the either Chi-squared value (Figure 5) if applicable or Fisher's exact test (Figure 6)

|  |
| --- |
| Fig 5. Chi-Squared P-Value for DiscoGEM Dataset at the Leaves Level with α = 0.3 |
|  |
| Note: 'fisher' here indicates that the P-value was calculated using Fisher's exact test. |

|  |
| --- |
| Fig 6. Fisher Exact Test P-Value for DiscoGEM Dataset at the Leaves Level with α = 0.3 |
|  |
| Note: ‘chi2’ here indicates that the P-value was calculated using Chi-Squared test. |

Figures 7 and 8 show the P-values for the DiscoGEM at level two with α = 0.3.

|  |
| --- |
| Fig 7. Chi-Squared P-Value for DiscoGEM Dataset at the Leaves-2 with α = 0.3 |
|  |
| Note: 'fisher' here indicates that the P-value was calculated using Fisher's exact test. |

|  |
| --- |
| Fig 8. Fisher Exact Test P-Value for DiscoGEM Dataset at Level-2 with α = 0.3 |
|  |
| Note: ‘chi2’ here indicates that the P-value was calculated using Chi-Squared test. |

A P-value less than 0.05 indicates an association between the sense pair. However, it does not specify whether this association is positive (the focus of this report) or negative. To determine the nature of the association, the Odds Ratio (OR) was calculated in the subsequent step.

Note that Chi-squared, Fisher’s exact test, and P-value for all the 12 datasets can be found in the '3\_results\binary\analysis\_value\_matrices\csv\_files' directory.

## Odd Ration (OR)

The Odds Ratio (OR) was calculated whenever the P-value was less than 0.05. An OR greater than 1 indicates positive associations, while an OR less than 1 indicates negative associations. Figure 9 displays the ORs for pairs with a P-value less than 0.05. To address issues such as zero counts or small cell frequencies in contingency tables, a 'continuity correction' was applied by adding 0.5 to help stabilize the calculations and prevent division by zero errors or extreme values in cases where one or more cells have low counts[[1]](#footnote-1).

Note that the sense pairs positively and negatively associated, based on odds ratio (OR) calculations for the remaining datasets, can be found in 3\_results/summary\_report

|  |
| --- |
| Fig 9. OR Values for DiscoGEM Dataset at the Leaves Level with α = 0.3 |
|  |
| Note: Cells in green indicate a positive correlation, while cells in red indicate a negative correlation |

Based on OR values, we can conclude the positively associated Sense pairs (Figure 10).

|  |
| --- |
| Fig 10. Positively Associated Sense Pairs Based on OR Values for DiscoGEM Dataset at the Leaves Level with α = 0.3 |
|  |

Figure 11 displays the OR values at level-2, while Figure 12 lists the positively associated sense pairs at that level

|  |
| --- |
| Fig 11. OR Values for DiscoGEM Dataset at Level-2 with α = 0.3 |
|  |

|  |
| --- |
| Fig 12. Positively Associated Sense Pairs Based on OR Values for DiscoGEM Dataset at Level-2 with α = 0.3 |
|  |

Thus, the first iteration of the decision pipeline is as shown in Figure 13.

|  |
| --- |
| Fig 13. Decision Pipeline, First Iteration (Based on OR) |
| A diagram of a flowchart  Description automatically generated |

Note that OR values for all the 12 datasets can be found in the '3\_results\binary\analysis\_value\_matrices\csv\_files' directory.

## Additional Measure, Yule’s Q

Yule's Q is a measure of association for 2x2 contingency tables, ranging from -1 to 1. A value of 1 indicates a perfect positive association, -1 indicates a perfect negative association, and 0 signifies no association between the variables. A value of 0.5 indicates a moderate positive association, while a value of -0.5 indicates a negative one.

[[[ INSERT FORUMALS HERE ]]]

Figure 14 depicts Yule’s Q values for the DiscoGEM Dataset at the Leaves Level with α = 0.3

|  |
| --- |
| Fig14. Yule’s Q Positive and Negative Sense Pair Associations Highlighted in Green and Red Respectively (for the DiscoGEM Dataset at the Leaves Level with α = 0.3) |
|  |

Note that 'NA' values indicate cases where the denominator of Yule’s Q was zero and cannot be calculated. Figure 15 lists the positively associated pairs, while Figure 16 shows the negatively associated ones.

|  |
| --- |
| Fig 15. Positively Associated Sense Pairs Based on Yule’s Q Values for the DiscoGEM Dataset at the Leaves Level with α = 0.3 |
|  |
| Fig 16. Negatively Associated Sense Pairs Based on Yule’s Q Values for the DiscoGEM Dataset at the Leaves Level with α = 0.3 |
|  |

Note that Yule’s Q values of 0.8 and above (-0.8 and below) denote 'strong' positive (negative) association, while a value of 0.5 (-0.5) denotes a 'moderate' positive (negative) association. Figures 17, 18, and 19 show the same analysis for Level-2.

|  |
| --- |
| Fig 17. Yule’s Q Positive and Negative Sense Pair Associations Highlighted in Green and Red Respectively (for the DiscoGEM Dataset at Level-2 with α = 0.3) |
|  |

|  |
| --- |
| Fig 18. Positively Associated Sense Pairs Based on Yule’s Q Values for the DiscoGEM Dataset at Level-2 with α = 0.3 |
|  |

|  |
| --- |
| Fig 19. Positively Associated Sense Pairs Based on Yule’s Q Values for the DiscoGEM Dataset at Level-2 with α = 0.3 |
|  |

Note Yule’s Q values for all the 12 datasets can be found in the '3\_results\binary\analysis\_value\_matrices\csv\_files' directory.

Also, note that the sense pairs positively and negatively associated, based on Yule’s Q calculations for the remaining datasets, can be found in 3\_results/summary\_report

## Accounting for Yule ’Q results in the OR results

OR and Yule’s Q results are somehow similar but not identical. Comparing the aforementioned results reveals overlapping for some sense pairs for the positive association case (Fig. 20).

|  |
| --- |
| Fig 20 |
| Positively associated pairs based on OR (for DiscoGEM, leaves level at 0.3) |
|  |
| Positively associated pairs based on Yule’s Q (for DiscoGEM, leaves level at 0.3) |
|  |
| Positively associated overlapping sense pairs |
| Succession | arg2-as-detail  Arg1-as-denier | arge2-as-denier  Similarity | conjunction |

For the purpose of this report, we will continue with the Chi-Squared/Fisher’s exact test OR-based positively associated pairs. We will add an annotation of (YQ+) just to indicate that certain sense pairs were also deemed positively associated based on Yule’s Q test. Thus, the collection of positively associated sense pairs becomes as shown in Fig. 21

|  |
| --- |
| Fig 21 |
| synchronous | precedence  succession | arg2-as-detail YQ+  result | arg1-as-detail  arg1-as-denier | arg2-as-denier YQ+  arg2-as-denier | contrast  similarity | conjunction YQ+  arg2-as-instance | arg2-as-detail |

Yule’s Q also yields a set of sense pairs that are negatively associated (Figure 16), which is the opposite and disproves what we are looking for. To account for Yule’s Q-based negative associations, this set of sense pairs should be excluded from the collection of positively associated pairs (figure 21). After examination, no sense pairs were excluded.

Similar analysis is done for Level-2 (Fig. 22)

|  |
| --- |
| Fig 22 |
| synchronous | asynchronous  concession | contrast  similarity | conjunction YQ+  instantiation | level-of-detail |

Based on the above-mentioned discussion, the decision pipeline is modified as seen in Fig. 22.

|  |
| --- |
| Fig 22. Decision Pipeline, Second Iteration (Based on OR & Yule’s Q) |
|  |

## Closer Look at Contingency Tables

So far, the statistical methods consider all four cells in the contingency table. However, upon closer examination of the contingency table, the following observations can be made:

* Vote-vote cells: These cells represent cases where there has been confusion about the sense (in a sense pair) that exists in a single instance. For α = 0.3, it reads as follows: six annotators at least have their votes split into two senses for the same instance; three annotators at least voted for sense one, and another three annotators at least voted for sense two. This type of sense pair is what this report aims to find with statistical significance.
* Vote-¬Vote cells: These cells represent cases where there hasn't been confusion about the sense (in a sense pair) that exists in a single instance. For α = 0.3, 3 or more annotators voted for sense one while two to none annotators voted for sense two. It can be said that the annotators were able to distinguish between these two senses for that specific instance. This is the opposite of what the report is trying to find.
* ¬Vote-¬Vote cells: This type of cell is similar to the Vote-¬Vote cells.
* ¬Vote-¬Vote cells: This cell does not provide information about the sense pair in hand; the annotators voted for other senses, indicating there could have been confusion or no confusion between another sense pair.

Thus, other methods that do not consider ¬Vote-¬Vote cell count should be considered.

## Pointwise Mutual Information PMI

point mutual information, is a measure of association. It compares the probability of two events occurring together to what this probability would be if the events were independent.

A black background with white text

Description automatically generated

We will calculate PMI on the vote-vote cell. PMI on the vote-vote cell does not take into account the "no vote-no vote" cell in the contingency table.

I am still not sure if/how positive PMI prove statistical significant positive association, Ill look into more,

Positive PMI: Indicates a higher-than-expected co-occurrence of events, suggesting a non-random association between them. Higher values signify a stronger association, but what is high enough?.

Zero PMI: Implies independence between events; their co-occurrence matches what would be expected by chance.

Negative PMI: Signifies a lower-than-expected co-occurrence of events, suggesting that they tend to occur separately or even repel each other.

For the purpose of this report, I’ll use the negative one to reject sense pairs and remove them from the positively associated sets

|  |
| --- |
| Fig 23. PMI on V-V Cells Values for DiscoGEM Dataset at Leaves Level with α = 0.3 |
|  |

|  |
| --- |
| Fig 24. Negatively Associated Sense Pairs Based on PMI Values for DiscoGEM Dataset at Leaves Level with α = 0.3 |
|  |

Comparing the PMI based negatively associated with the positively associaeted pairs of figure 17, we notice two overlaps pairs (result | arg1-as-detail, arg2-as-denier| contrast, and arg2-as-instance | arg2-as-detail ). We exclude these pairs and we end up with the suriving postiviel associated sense pair fig XX

|  |
| --- |
| Fig. 25: Survived Positively Associated Sense Pair for DiscoGEM Dataset at Leaves Level with α = 0.3 |
| synchronous | precedence  succession | arg2-as-detail YQ+  arg1-as-denier | arg2-as-denier YQ+  similarity | conjunction YQ+ |

Similar analysis is done for Level-2 and yielded the results in Figure 26. (The pair 'instantiation | level-of-detail' was excluded).

|  |
| --- |
| Fig 26. Survived Positively Associated Sense Pair for DiscoGEM Dataset atLevel-2 with α = 0.3 |
| synchronous | asynchronous  concession | contrast  similarity | conjunction YQ+ |

Note PMI (on V-V cells) values for all the 12 datasets can be found in the '3\_results\binary\analysis\_value\_matrices\csv\_files' directory.

Also, note that the sense pairs negatively associated, based on PMI on V-V cell calculations for the remaining datasets, can be found in 3\_results/summary\_report

Based on the above-mentioned discussion, the decision pipeline is modified as seen in Fig. 27.

## Decision Making Pipeline

The decision pipeline for generating the 'survived' positively associated pairs begins with transforming raw data into binary form based on different thresholds. This facilitates pairwise comparisons. Contingency tables for each pair are then constructed, and statistical tests (Chi-squared, Fisher's exact test) calculate p-values to identify significant associations. Odd Ratio (OR) values discern the direction of these associations, with OR>1 indicating a positive relationship. Yule's Q measure further assesses the strength and direction, with positive associations annotated as YQ+ and negative overlap actively removed. Finally, the analysis incorporates pointwise mutual information (PMI) on the Vote-Vote cells to refine identifying positively associated sense pairs by excluding pairs with overlapping negative associations identified through PMI analysis.

the pipeline is not fully automated; it reaches up to OR, and calculates Yule’s Q and PMI independently, and I manually annotate and exclude pairs. If you are satisfied with the pipeline and/or would like adjustments, I can modify the code to calculate the 'survivor pairs' for the remaining 10 datasets. If so, please add that to the to do list on the first page

|  |
| --- |
| Fig 27. Decision Pipeline, Third and Final Iteration (Based on OR, Yule’s Q and PMI) |
| A diagram of a computer program  Description automatically generated with medium confidence |

# Discussing the Findings

Analyzing the contingency tables for the four sense pairs that showed positive association in the DiscoGEM dataset at the Leaves Level, after converting to binary data with a 0.3 threshold, we observe the following

|  |  |
| --- | --- |
| Fig 28. |  |
| precedence V ¬V  synchronous  V 13 49  ¬V 589 5854 | arg2-as-detail V ¬V  succession  V 6 6  ¬V 1410 5083 |
| arg2-as-denier V ¬V  arg1-as-denier  V 23 98  ¬V 330 6054 | conjunction V ¬V  similarity  V 60 20  ¬V 2247 4178 |

In all cases, the majority of instances fall into the ¬V-¬V category, where annotators consistently did not choose a particular sense.

The V-V category, indicating split decisions among annotators, has the fewest or nearly the fewest counts, but it's crucial for understanding annotator confusion. Consider the conjunction | similarity pair. Out of 2,327 (60 + 20 + 2247) instances where they appear in our dataset:

* In 60 cases, annotators were split between these senses.
* In 2,267 (2247 + 20) cases, a consensus favored one sense over the other.
* This suggests a 2.6% (60 / 2327) confusion rate among annotators when presented with these pairs, with a clear preference (no confusion) established in 97.4% of cases.

The V-V counts are particularly informative because they directly reflect instances of annotator confusion. To quantify the extent of annotator confusion for each sense pair, an indicator formula is suggested:

Proposed indicator

The proposed indicator's value spans from zero, indicating no confusion among annotators (where count(V,V) = 0), to one, highlighting complete indecision (where both count(V,¬V) and count(¬V,V) equal zero). This metric effectively captures the proportion of instances in which a significant number of annotators (in this context, more than six) were split in their decisions between sense pairs, thereby providing a precise measure to assess levels of confusion.

Figures 29 and 30 show the proposed indicator values for the DiscoGEM dataset at the Leaves and Level-2 with α = 0.3.

Notice at the Leaves Level, the maximum value of the proposed indicator is 10.8% which offer a different explanation to the annotator behavior challenges previous findings by quantifying the extent of confusion in sense selection.

|  |
| --- |
| Figure 29: Proposed Indicator Values for the DiscoGEM Dataset at the Leaves Level with α = 0.3 |
| A graph with numbers and a number of words  Description automatically generated with medium confidence |

|  |
| --- |
| Figure 29: Proposed Indicator Values for the DiscoGEM Dataset at the Level 2 with α = 0.3 |
| A screenshot of a graph |

# Appendices

## Appendix A: Calculation Verification Sheet

In this sheet, you can choose the corpus, the alpha threshold for binary transformation, and the sense pair that you want to verify the calculation for. Note that Fisher Exact’s test is not calculated in that sheet. Also, currently it only accounts for the leaves level.

If any of the ‘Expected Values’ Table cells is less than 5, that cell is highlighted, indicating that the Chi2 and P-value (although calculated in that sheet) should not be taken into consideration. For that end, we used Fisher Exact’s test. This is implemented in the code but not in the value verification sheet.

Additionally, you can navigate various sheets in the Excel file to check how each method was calculated. Note that this sheet is independent of the code base and crafted manually. However, it uses the same calculation logic used in the code. The values generated from this sheet are used for verification purposes only. If you want me to update this sheet so it includes level-2 too, please indicate that in the to-do table.

A screenshot of a table

Description automatically generated

## Appendix B: Overview of Code-Generated Files and Directory Structure

* 0\_raw\_data: Contains the CSV files of the DiscoGEM and QADC datasets as they are (2 datasets).
* 1\_ready\_to\_transform: Contains the cleaned and grouped data before the transformation to binary (4 datasets). See details in section 3.a.
* 2\_ready\_to\_process: Contains the data transformed into binary format (12 datasets). For more details, refer to section 3.b.
* 3\_results: Contains multiple folders:
  + Summary\_report: Contains the results for each statistical method used, which, in turn, contains the results for the 12 datasets with the positive and negative associations.
  + Binary: Contains multiple folders:
    - Contingency\_tables: For the 12 datasets.
    - Expected\_tables: For the 12 datasets.
    - Analysis\_value\_matrices: Contains a value matrix for each statistical method for each dataset.

1. Add Resource [↑](#footnote-ref-1)