COMP 490: COMPUTER SCIENCE PROJECT I

Statistical Analysis of Semantic Sense Pairs in Discourse Annotations

Supervisor: Professor Leila Kosseim

Course Coordinator: Dr. David K. Probst

Nawar Turk

Concordia University

April 15th, 2024

"I certify that this submission is my original work and meets the Faculty's Expectations of Originality"

Contents

Table of Figures

Executive Summary

The report aims to identify commonly confused sense pairs in semantic annotations based on the PDTB-3 Sense Hierarchy levels 2 and 3. The annotation data was converted to a binary format to analyze the sense associations statistically. To determine the associations' nature, a positively associated sense pair decision pipeline was created, calculating P-values and Odds Ratios using the Chi-squared Test and Fisher's Exact tests. Pointwise Mutual Information (PMI) was also used in the pipeline. An additional indicator was proposed to analyze the decision pipeline's sanity.

The results show some sense pairs with statistically significant positive correlations. Still, it's advisable to interpret the results cautiously due to the 'no vote, no vote' count's high incidence, possibly driving these correlations, which is not meaningful considering the report's objective.

The report suggests that the positively associated sense pairs identified may be influenced by the annotation experiment's structure, calling for more focused analyses on 'vote, vote' cells that should accurately reflect annotator confusion.

1. Objective

This report aims to identify commonly confused pairs of semantic senses, particularly at levels 2 and 3 of the PDTB-3 Sense Hierarchy[1], using data from the DiscoGeM and QADC datasets. DiscoGeM[2] is a crowdsourced corpus containing 6,505 instances of inter-sentential implicit discourse relations gathered from three genres: political speeches, literature, and encyclopedic texts. Ten crowd workers annotated each instance.

Figure 1

PDTB-3 Sense Hierarchy Table

Level-1	Level-2	Level-3
TEMPORAL	SYNCHRONOUS	–
	ASYNCHRONOUS	PRECEDENCE
		SUCCESSION
CONTINGENCY	CAUSE	REASON
		RESULT
		NEGRESULT
	CAUSE+BELIEF	REASON+BELIEF
		RESULT+BELIEF
	CAUSE+SPEECHACT	REASON+SPEECHACT
		RESULT+SPEECHACT
	CONDITION	ARG1-AS-COND
		ARG2-AS-COND
	CONDITION+SPEECHACT	–
	NEGATIVE-CONDITION	ARG1-AS-NEGCOND
		ARG2-AS-NEGCOND
	NEGATIVE-CONDITION+SPEECHACT	–
	PURPOSE	ARG1-AS-GOAL
		ARG2-AS-GOAL
COMPARISON	CONCESSION	ARG1-AS-DENIER
		ARG2-AS-DENIER
	CONCESSION+SPEECHACT	ARG2-AS-DENIER+SPEECHACT
	CONTRAST	–
	SIMILARITY	–
EXPANSION	CONJUNCTION	–
	DISJUNCTION	–
	EQUIVALENCE	–
	EXCEPTION	ARG1-AS-EXCPT
		ARG2-AS-EXCPT
	INSTANTIATION	ARG1-AS-INSTANCE
		ARG2-AS-INSTANCE
	LEVEL-OF-DETAIL	ARG1-AS-DETAIL
		ARG2-AS-DETAIL
	MANNER	ARG1-AS-MANNER
		ARG2-AS-MANNER
	SUBSTITUTION	ARG1-AS-SUBST
		ARG2-AS-SUBST

[1] Rashmi Prasad, Bonnie Webber, Alan Lee, Aravind Joshi. The Penn Discourse Treebank 3.0 Annotation Manual. Philadelphia: Linguistic Data Consortium, 2019. LDC Catalog No.: LDC2019T05. Web. March 15, 2019.

[2] Scholman, M., Dong, T., Yung, F., & Demberg, V. (2022, June 1). DiscoGeM: A Crowdsourced Corpus of Genre-Mixed Implicit Discourse Relations (N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Marian

2. Data Processing Workflow

a. Raw Data Cleaning and Hierarchical Grouping

The data we received for analysis had information on the annotator votes at the leaf level. The raw data can be found in this directory 0_raw_data. We removed unnecessary columns not mentioned in the Sense Hierarchy table (Figure 1), resulting in 27 sense columns for DiscoGeM and 28 sense columns for QADC. We grouped the votes for senses that constitute level 2 together to form new columns labelled appropriately based on the PDTB-3 Sense Hierarchy table to form level 2 columns. As a result, we have four datasets stored under the 1_ready_to_transform directory.

b. 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 Manning and Schütze's Foundations of Statistical Natural Language Processing, specifically as described in Chapter 5 on 'Collocations'[3]. To this end, different threshold values (α) were used for sensitivity analysis (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 three or more out of the ten annotators choose a certain sense for a certain instance, we consider this sense to be voted for. Otherwise (two annotators or less), we ignored these votes.

Thus, twelve datasets were created (4 datasets * 3 thresholds = 12 datasets). Figure 2 displays the twelve datasets to be processed. In this report, we will show details of the results of only the two highlighted datasets; these are DiscoGeM with a threshold of 0.3 at both levels. As for all other datasets, only the final results are presented. The binary transformed datasets are stored in the 2_ready_to_process/binary directory.

c. Constructing Contingency Tables

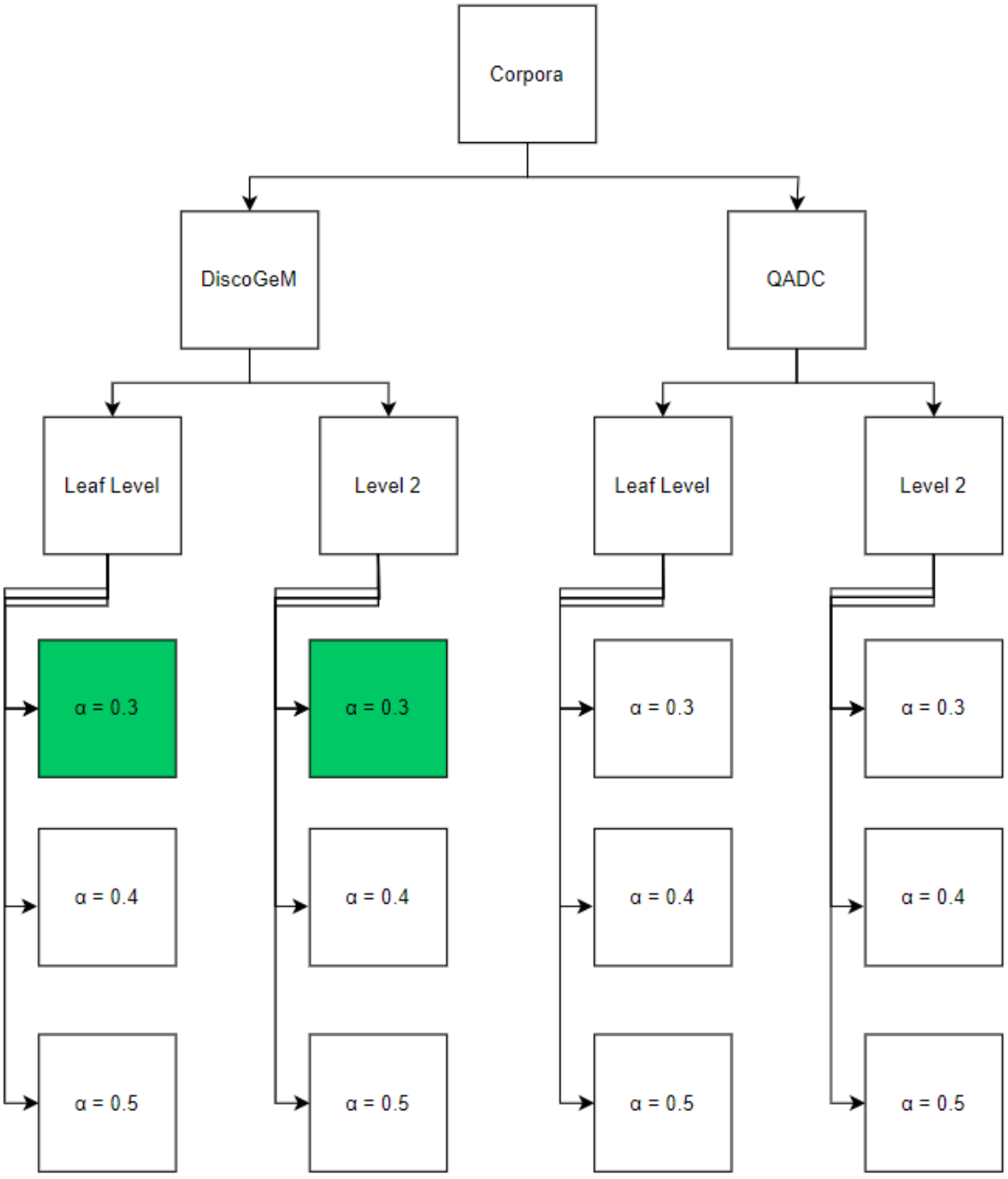
Doing a pairwise comparison between each possible sense pair, we construct a 2x2 contingency table with four cells (Figure 3). Where:

- 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.

[3] Manning, Christopher D., and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. 4th printing. Cambridge, MA: MIT Press, 2001. Chapter 5, "Collocations."

Figure 2

Twelve Datasets Processed Following Binary Transformation



Note: Detailed calculations are shown only for the datasets highlighted in green. Only the results are presented for all other datasets.

Figure 3

Contingency Table Structure

	Sense 2	
Sense 1	V	¬V
V	a	b
¬V	c	d

All contingency tables are in the 3_results/binary/contingency_tables directory. Note that during the transformation process, any columns that only had ¬V values (no vote) were dropped from further analysis. This resulted in each row and column of the contingency table having at most one cell with a count of zero. Figure 4 shows a sample of the contingency tables considered during the processing phase.

Figure 4

Examples of Contingency Tables After Transformation

```
77
78    -.-.-.-.-.-.-.-.-.-
79    arg2-as-detail     V    ¬V
80    synchronous
81    V                  4    58
82    ¬V              1412  5031
83
84    -.-.-.-.-.-.-.-.-.-
85    arg2-as-subst   V    ¬V
86    synchronous
87    V               0    62
88    ¬V             14  6429
89
90    -.-.-.-.-.-.-.-.-.-
91    succession   V    ¬V
92    precedence
93    V            0   602
94    ¬V          12  5891
95
```

3. Data Analysis

a. Chi-squared Test, Fisher's Exact Test and P-Value

The Chi-squared Test values were calculated for all eligible pairs. When calculating the Chi-squared Test value, the expected table is constructed. As recommended by the literature[4], all cells of the expected table should have a value greater than 5 for the Chi-squared Test calculation to be reliable. However, if any cell of the expected table contains a value less than five, 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 (Figure 5). This was feasible due to the 2X2 dimensions of the contingency table[3].

Figure 5

Example of an Expected Table with Cell Values Less Than Five

contrast	V	¬V
synchronous		
V	1	60
¬V	174	6268

Subsequently, the P-values for DiscoGeM at leaf level with $\alpha = 0.3$ are calculated from either the Chi-squared Test value, if applicable or Fisher's Exact Test (Figure 6). Similarly, Figure 7 shows the P-values for the DiscoGeM at level 2 with $\alpha = 0.3$.

A P-value less than 0.05[5] indicates an association between the sense pair. However, it does not specify whether this association is positive (the focus of this report) or negative. The odds ratio (OR) was calculated in the subsequent step to determine the nature of the association.

The 3_results\binary\analysis_value_matrices\csv_files directory contains the Chi-squared, Fisher's exact test, and P-value for all 12 datasets.

[4] Nowacki, Amy. "Chi-square and Fisher's exact tests." *Cleveland Clinic Journal of Medicine*, vol. 84, no. 9 Suppl 2, Sep. 2017, pp. e20-e25. DOI: 10.3949/ccjm.84.s2.04. Accessed 13 Apr. 2024. https://www.ccjm.org/content/84/9_suppl_2/e20.

[5] Lumen Learning. "Introduction to Statistics Corequisite." Accessed April 13, 2024. https://courses.lumenlearning.com/introstats1/.

Figure 6

Chi-squared Test and Fisher's Exact Test P-Values for DiscoGeM Dataset at Leaf Level with α = 0.3

	synchronous	precedence	succession	reason	result	arg1-as-goal	arg1-as-denier	arg2-as-denier	contrast	similarity	conjunction	disjunction	arg2-as-instance	arg1-as-detail	arg2-as-detail	arg2-as-subst
synchronous		0.00	fisher	fisher	0.17	fisher	fisher	fisher	fisher	fisher	0.79	fisher	fisher	fisher	0.00	fisher
precedence			fisher	0.00	0.00	fisher	0.00	0.00	0.00	0.00	0.00	fisher	0.00	0.02	0.00	fisher
succession				fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher
reason					0.00	fisher	0.02	0.00	0.03	0.01	0.00	fisher	0.01	fisher	0.41	fisher
result						fisher	0.00	0.00	0.00	0.00	0.00	fisher	0.00	0.03	0.00	fisher
arg1-as-goal							fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher
arg1-as-denier								0.00	fisher	fisher	0.00	fisher	0.01	fisher	0.00	fisher
arg2-as-denier									0.03	fisher	0.00	fisher	0.00	fisher	0.00	fisher
contrast										fisher	0.00	fisher	0.00	fisher	0.00	fisher
similarity											0.00	fisher	fisher	fisher	0.00	fisher
conjunction												fisher	0.00	0.00	0.00	fisher
disjunction													fisher	fisher	fisher	fisher
arg2-as-instance														fisher	0.00	fisher
arg1-as-detail															0.13	fisher
arg2-as-detail																fisher
arg2-as-subst																

	synchronous	precedence	succession	reason	result	arg1-as-goal	arg1-as-denier	arg2-as-denier	contrast	similarity	conjunction	disjunction	arg2-as-instance	arg1-as-detail	arg2-as-detail	arg2-as-subst
synchronous		chi2	1.00	0.02	chi2	1.00	0.63	0.05	0.42	1.00	chi2	1.00	0.05	1.00	chi2	1.00
precedence			0.62	chi2	chi2	1.00	chi2	chi2	chi2	chi2	chi2	1.00	chi2	chi2	chi2	1.00
succession				1.00	0.01	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.52	1.00	0.03	1.00
reason					chi2	1.00	chi2	chi2	chi2	chi2	chi2	1.00	chi2	1.00	chi2	1.00
result						0.32	chi2	chi2	chi2	chi2	chi2	1.00	chi2	chi2	chi2	0.57
arg1-as-goal							1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
arg1-as-denier								chi2	0.26	0.41	chi2	1.00	chi2	0.64	chi2	1.00
arg2-as-denier									chi2	0.02	chi2	1.00	chi2	0.59	chi2	1.00
contrast										0.73	chi2	1.00	chi2	0.27	chi2	0.32
similarity											chi2	1.00	0.63	0.58	chi2	1.00
conjunction												0.35	chi2	chi2	chi2	0.00
disjunction													1.00	1.00	1.00	1.00
arg2-as-instance														0.44	chi2	1.00
arg1-as-detail															chi2	1.00
arg2-as-detail																1.00
arg2-as-subst																

Note: 'fisher' indicates that the P-value was calculated using Fisher's Exact Test. 'chi2' indicates that the P-value was calculated using the Chi-squared Test. Cells highlighted in green represent pairs with a P-value less than 5%.

Figure 7

Chi-squared Test and Fisher's Exact Test P-Values for DiscoGeM Dataset at Level 2 with α = 0.3

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		synchronous	asynchronous	cause	purpose	concession	contrast	similarity	conjunction	disjunction	instantiation	level-of-detail	substitution
2	synchronous		0.00	0.00	fisher	0.01	fisher	fisher	0.79	fisher	fisher	0.00	fisher
3	asynchronous			0.00	fisher	0.00	0.00	0.00	0.00	fisher	0.00	0.00	fisher
4	cause				fisher	0.00	0.00	0.00	0.00	fisher	0.00	0.00	0.53
5	purpose					fisher	fisher	fisher	fisher	fisher	fisher	fisher	fisher
6	concession						0.00	0.01	0.00	fisher	0.00	0.00	fisher
7	contrast							fisher	0.00	fisher	0.00	0.00	fisher
8	similarity								0.00	fisher	fisher	0.00	fisher
9	conjunction									fisher	0.00	0.00	fisher
10	disjunction										fisher	fisher	fisher
11	instantiation											0.00	fisher
12	level-of-detail												fisher
13	substitution												

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		synchronous	asynchronous	cause	purpose	concession	contrast	similarity	conjunction	disjunction	instantiation	level-of-detail	substitution
2	synchronous		chi2	chi2	1.00	chi2	0.42	1.00	chi2	1.00	0.03	chi2	1.00
3	asynchronous			chi2	1.00	chi2	chi2	chi2	chi2	1.00	chi2	chi2	1.00
4	cause				0.44	chi2	chi2	chi2	chi2	1.00	chi2	chi2	chi2
5	purpose					1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6	concession						chi2	chi2	chi2	1.00	chi2	chi2	0.63
7	contrast							0.73	chi2	1.00	chi2	chi2	0.32
8	similarity								chi2	1.00	0.49	chi2	1.00
9	conjunction									0.35	chi2	chi2	0.00
10	disjunction										1.00	1.00	1.00
11	instantiation											chi2	1.00
12	level-of-detail												1.00
13	substitution												

Note: 'fisher' indicates that the P-value was calculated using Fisher's Exact Test. 'chi2' indicates that the P-value was calculated using the Chi-squared Test. Cells highlighted in green represent pairs with a P-value less than 5%.

b. Odds Ratio (OR)

The P-value indicates whether there is a statistically significant association between variables (a sense pair in our case). However, it does not provide information about the direction of this association.

To achieve the report's objective of identifying commonly confused pairs of senses, we need to find sense pairs in which if one sense gets a vote, the other sense in the same pair also gets a vote for the same instance. This indicates that the annotators did not agree on the sense label for a single instance. Consequently, the two paired senses should exhibit similar behaviour, and thus, the direction of the association should be positive.

This is why the Odds Ratio (OR) for the pairs with P-value < 0.05 was calculated. An OR greater than 1 suggests positive associations, while an OR less than 1 indicates negative associations [6]. To avoid the problem of having zero counts in the contingency table, which could result in division by zero when calculating OR, a continuity correction[7] was applied by adding 0.5 to the four cells in the contingency table.

Figure 8 shows the ORs for pairs with a P-value less than 0.05 for the leaf level and the list of positively associated pairs (OR > 1). Figure 9 displays the OR values at level 2 and lists the positively associated sense pairs.

Note that the pairs of senses that are positively and negatively associated, based on the P-value and the odds ratio (OR) calculations for the remaining datasets, can be found in this directory 3_results/summary report

[6] "Odds Ratio," Statistics Easily, accessed April 13, 2024, https://statisticseasily.com/odds-ratio/.
[7] Wikipedia contributors. "Yates's correction for continuity." Wikipedia, The Free Encyclopedia. Last modified February 10, 2024. https://en.wikipedia.org/wiki/Yates%27s_correction_for_continuity.

Figure 8

OR Values and Positively Associated Sense Pairs Based on OR for DiscoGeM Dataset at Leaf Level with α = 0.3

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		synchronous	precedence	succession	reason	result	arg1-as-goal	arg1-as-denier	arg2-as-denier	contrast	similarity	conjunction	disjunction	arg2-as-instance	arg1-as-detail	arg2-as-detail	arg2-as-subst
2	**synchronous**		2.7	NA	0.1	NA	NA	NA	0.1	NA	NA	NA	NA	NA	NA	0.3	NA
3	**precedence**			NA	0.1	0.8	NA	0.1	0.5	0.1	0.1	0.4	NA	0.1	0.2	0.1	NA
4	**succession**				NA	0.1	NA	NA	NA	NA	NA	NA	NA	NA	NA	3.6	NA
5	**reason**					0.2	NA	0.3	0.1	0.4	0.1	0.3	NA	0.5	NA	NA	NA
6	**result**						NA	0.3	0.5	0.2	0.2	0.3	NA	0.2	1.7	0.2	NA
7	**arg1-as-goal**							NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	**arg1-as-denier**								4.4	NA	NA	0.1	NA	0.1	NA	0.4	NA
9	**arg2-as-denier**									1.8	0.1	0.3	NA	0.2	NA	0.1	NA
10	**contrast**										NA	0.3	NA	0.1	NA	0.2	NA
11	**similarity**											5.5	NA	NA	NA	0.2	NA
12	**conjunction**												NA	0.3	0.2	0.6	0.1
13	**disjunction**													NA	NA	NA	NA
14	**arg2-as-instance**														NA	1.7	NA
15	**arg1-as-detail**															NA	NA
16	**arg2-as-detail**																NA
17	**arg2-as-subst**																

```
*** discogem_multi_leaves_0.3 ***

 Positive Association
synchronous          | precedence           | 2.7085  |
succession           | arg2-as-detail       | 3.604   |
result               | arg1-as-detail       | 1.679   |
arg1-as-denier       | arg2-as-denier       | 4.3706  |
arg2-as-denier       | contrast             | 1.8253  |
similarity           | conjunction          | 5.4868  |
arg2-as-instance     | arg2-as-detail       | 1.7317  |
_____________________________________
```

Note. Cells in green indicate a positive correlation (OR > 1), while cells in red indicate a negative correlation (OR < 1). 'NA' indicates that the sense pair has a P-value greater than 0.05, and thus, the OR was not calculated due to the lack of a statistically significant association to investigate.

Figure 9

OR Values and the Resulted Positively Associated Sense Pairs for DiscoGeM Dataset at Level 2 with α = 0.3

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		synchronous	asynchronous	cause	purpose	concession	contrast	similarity	conjunction	disjunction	instantiation	level-of-detail	substitution
2	**synchronous**		2.5	0.4	NA	0.1	NA	NA	NA	NA	0.1	0.2	NA
3	**asynchronous**			0.5	NA	0.3	0.1	0.1	0.5	NA	0.1	0.1	NA
4	**cause**				NA	0.3	0.2	0.1	0.3	NA	0.3	0.4	NA
5	**purpose**					NA	NA	NA	NA	NA	NA	NA	NA
6	**concession**						2.4	0.2	0.3	NA	0.2	0.3	NA
7	**contrast**							NA	0.3	NA	0.1	0.2	NA
8	**similarity**								5.5	NA	NA	0.2	NA
9	**conjunction**									NA	0.3	0.5	0.1
10	**disjunction**										NA	NA	NA
11	**instantiation**											1.5	NA
12	**level-of-detail**												NA
13	**substitution**												

```
*** discogem_multi_level2_0.3 ***

 Positive Association
synchronous          | asynchronous          | 2.4841  |
concession           | contrast              | 2.3523  |
similarity           | conjunction           | 5.4868  |
instantiation        | level-of-detail       | 1.5384  |
_____________________________________
```

Note. Cells in green indicate a positive correlation, while cells in red indicate a negative correlation. 'NA' indicates that the sense pair has a P-value greater than 0.05, and thus the OR was not calculated due to the lack of a statistically significant association to investigate.

Thus, the first iteration of a decision pipeline to identify positively associated sense pairs has been developed and is shown in Figure 10. To simplify the pipeline, only Fisher calculations were used for all 12 datasets instead of combining the Chi-squared Test and Fisher. The results obtained through this simplified pipeline were comparable to those obtained through the previous pipeline. Therefore, we have updated the pipeline to use Fisher calculations exclusively. Figure 11 displays the simplified decision pipeline.

Figure 10

Decision Pipeline for Positively Associated Sense Pairs (First Iteration)

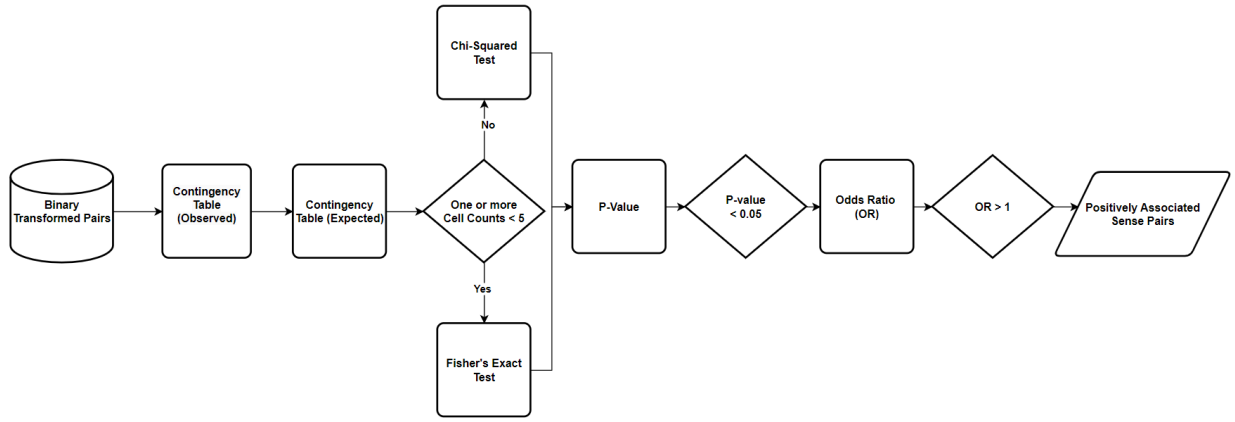
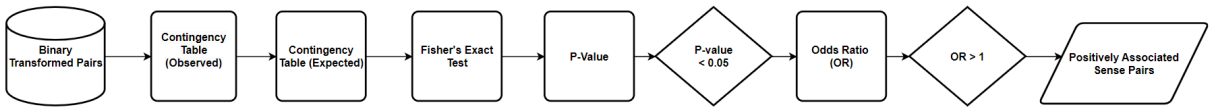


Figure 11

Simplified Decision Pipeline for Positive Associations (Second Iteration)



c. A Closer Look at Contingency Tables

The 'vote, no vote' cells in the table indicate that annotators were not confused about which sense to choose for a particular instance. For example, when alpha (α) is 0.3, at least three annotators chose the first sense, and no more than two chose the second sense. This indicates that the annotators were able to distinguish between the two senses accurately for that instance. However, this aspect is not the focus of our report.

Similarly, the 'no vote, vote' cells in the table indicate accurate sense differentiation by the annotators. The 'no vote, no vote' cells do not provide any information about the sense pair in question, as the annotators voted for other senses, making it unclear whether there was confusion between the sense pair constituting the contingency table in hand.

On the other hand, the ‘vote, vote’ cells represent uncertainty about the most appropriate sense for a specific instance. This is because these cells indicate that at least six annotators, when $\alpha = 0.3$, have split their votes between the two senses of the pair for the same instance, with three or more voting for each sense. Therefore, the report suggests that these are the cells that one should

focus on to achieve the report's goal. The Pointwise Mutual Information (PMI) method has analyzed these 'vote, vote' cells.

d. Pointwise Mutual Information PMI

Pointwise Mutual Information (PMI) is a measure of association that compares the probability of two events occurring together to the probability of the events being independent[2]. In our case, we are interested in calculating PMI on the 'vote, vote' cells. Positive PMI indicates a higher-than-expected co-occurrence of events, suggesting a non-random association between them. Negative PMI signifies a lower-than-expected co-occurrence of events, suggesting that they tend to occur separately or even repel each other. At the same time, zero or close to zero PMI implies independence between events.

To find sense pairs with a positive association, we aim to identify higher PMI values; however, it can be challenging to determine what is considered high enough. For our report, we will use a certain value of PMI as a threshold below which we will consider the sense pairs to have either negative or no association. This is because the literature[2] suggests that the PMI measure is a good measure for independence (no association) and a bad measure for dependence (associations), and we will use it as a tool to reject pairs at the end of the pipeline if they have a value less than a certain limit. While it is unclear what the value of this limit should be, examining the contingency table presented in the literature[2] suggests that a value of 1 could be used.

To summarize, if the PMI value is less than one, we will reject the positive association of the pairs initially deemed positively associated based on the OR and P-value. Consequently, we will denote such pairs as 'PMI-rejected' sense pairs. Figure 12 shows PMI Values for 'vote, vote' cells in the DiscoGeM dataset at leaf level with $\alpha = 0.3$.

After comparing the PMI-rejected sense pairs with the positively associated pairs of Figure 8, we found that three pairs overlap (result & arg1-as-detail, arg2-as-denier & contrast, and arg2-as-instance & arg2-as-detail). To get the 'survived positively associated sense pair' (figure 13), we removed these pairs. We did a similar analysis for level 2, which yielded results in Figure 14, where the pair of instantiation & level-of-detail was excluded.

It's worth noting that the PMI on 'vote, vote' cell values for all 12 datasets can be found in the 3_results\binary\analysis_value_matrices\csv_files directory. Additionally, a summary of the PMI-rejected sense pair can be found in 3_results/summary_report. Based on the above discussion, we modified the decision pipeline, and the changes can be seen in Figure 15.

Figure 12

PMI Values for 'Vote, Vote' Cells in the DiscoGeM Dataset at Leaf Level with α = 0.3 and the List of PMI-Rejected Sense Pairs

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		**synchronous**	**precedence**	**succession**	**reason**	**result**	**arg1-as-goal**	**arg1-as-denier**	**arg2-as-denier**	**contrast**	**similarity**	**conjunction**	**disjunction**	**arg2-as-instance**	**arg1-as-detail**	**arg2-as-detail**	**arg2-as-subst**
2	**synchronous**		1.2	NA	NA	-0.4	NA	NA	NA	NA	NA	0.1	NA	NA	NA	-1.8	NA
3	**precedence**			NA	-4.4	-0.3	NA	-3.5	-0.9	-3.0	NA	-0.8	NA	-3.6	-2.7	-3.4	-0.4
4	**succession**				NA	NA	NA	NA	NA	NA	NA	-0.1	NA	0.5	NA	1.2	NA
5	**reason**					-2.0	NA	-2.1	-3.0	-1.3	NA	-1.3	NA	-0.9	-0.3	-0.1	0.0
6	**result**						1.6	-1.2	-0.8	-1.7	-2.1	-0.8	NA	-2.0	0.5	-1.6	-0.6
7	**arg1-as-goal**							NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	**arg1-as-denier**								1.8	0.6	NA	-2.4	NA	NA	NA	-1.1	NA
9	**arg2-as-denier**									0.7	NA	-1.4	NA	-2.1	-0.9	-2.6	NA
10	**contrast**										-1.1	-1.3	NA	-3.4	NA	-2.1	1.4
11	**similarity**											1.1	NA	-0.6	0.2	-2.5	NA
12	**conjunction**												1.5	-1.2	-2.3	-0.5	NA
13	**disjunction**													NA	NA	NA	NA
14	**arg2-as-instance**														-1.0	0.5	NA
15	**arg1-as-detail**															-0.6	NA
16	**arg2-as-detail**																0.0
17	**arg2-as-subst**																

```
*** discogem_multi_leaves_0.3 ***

synchronous          | result              | -0.4182 |
synchronous          | conjunction         | 0.0649  |
precedence           | result              | -0.2647 |
precedence           | arg2-as-denier      | -0.9423 |
precedence           | conjunction         | -0.8074 |
precedence           | arg2-as-subst       | -0.3736 |
succession           | conjunction         | -0.0894 |
succession           | arg2-as-instance    | 0.5049  |
reason               | arg2-as-instance    | -0.9165 |
reason               | arg1-as-detail      | -0.2757 |
reason               | arg2-as-detail      | -0.1063 |
reason               | arg2-as-subst       | 0.0462  |
result               | arg2-as-denier      | -0.79   |
result               | conjunction         | -0.7912 |
result               | arg1-as-detail      | 0.454   |
result               | arg2-as-subst       | -0.5933 |
arg1-as-denier       | contrast            | 0.611   |
arg2-as-denier       | contrast            | 0.7444  |
arg2-as-denier       | arg1-as-detail      | -0.9255 |
similarity           | arg2-as-instance    | -0.6471 |
similarity           | arg1-as-detail      | 0.2161  |
conjunction          | arg2-as-detail      | -0.4802 |
arg2-as-instance     | arg2-as-detail      | 0.5412  |
arg1-as-detail       | arg2-as-detail      | -0.6076 |
arg2-as-detail       | arg2-as-subst       | -0.0227 |
```

Note: Cells highlighted in red represent pairs with PMI values less than 1

Figure 13

Survived Positively Associated Sense Pairs for DiscoGeM Dataset at Leaf Level with α = 0.3

synchronous	precedence
succession	arg2-as-detail
arg1-as-denier	arg2-as-denier
similarity	conjunction

Figure 14

Survived Positively Associated Sense Pair for DiscoGeM Dataset at Level 2 with α = 0.3

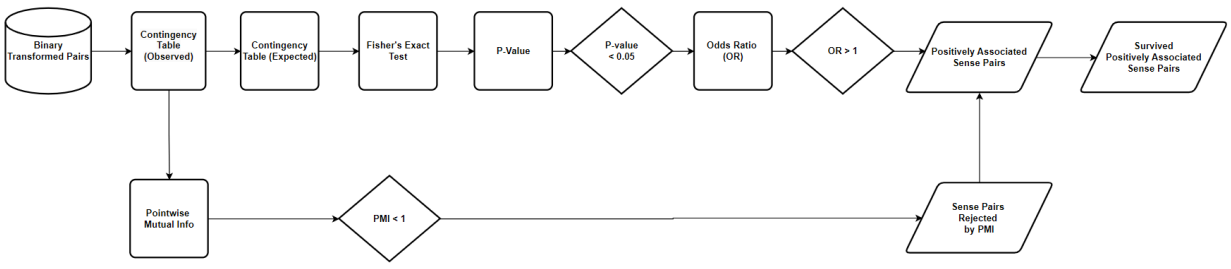
synchronous	asynchronous
concession	contrast
similarity	conjunction

e. Decision Pipeline for Positively Associated Sense Pairs

The decision pipeline (Figure 15) 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. Next, by using Fisher's Exact Test, we calculate P-values to identify significant associations. Odds Ratio (OR) values are used next to discern the direction of these associations, with OR>1 indicating a positive relationship. Finally, the analysis incorporates pointwise mutual information (PMI) on the 'vote, vote' cells to refine and identify positively associated sense pairs (denoted as survived positively associated pairs) by excluding pairs with PMI values less than one.

Figure 15

Final Decision Pipeline for Positive Associations (Third Iteration)



4. Results

Figures 16 and 17 show the results obtained from the detailed analyses described in this report, in addition to the final results for the rest of the datasets.

Figure 16

Survived Positively Associated Sense Pairs for All Datasets at Leaf Level

Corpus	DiscoGeM Dataset	QADC Dataset
Leaf Level, Threshold of 0.3		
Initial positively associated sense pairs	synchronous, precedence succession, arg2-as-detail result, arg1-as-detail arg1-as-denier, arg2-as-denier arg2-as-denier, contrast similarity, conjunction arg2-as-instance, arg2-as-detail	precedence, succession arg1-as-denier, contrast
Survived positively associated sense pairs	synchronous, precedence succession, arg2-as-detail arg1-as-denier, arg2-as-denier similarity, conjunction	precedence, succession arg1-as-denier, contrast
Leaf Level, Threshold of 0.4		
Initial positively associated sense pairs	result, arg1-as-detail similarity, conjunction	None
Survived positively associated sense pairs	result, arg1-as-detail similarity, conjunction	None
No positively associated pairs were found for either dataset at leaf level with a threshold of 0.5		

Figure 17

Survived Positively Associated Sense Pairs for All Datasets at Level 2

Corpus	DiscoGeM Dataset	QADC Dataset
Level 2, Threshold of 0.3		
Initial positively associated sense pairs	synchronous, asynchronous concession, contrast similarity, conjunction instantiation, level-of-detail	concession, contrast
Survived positively associated sense pairs	synchronous, asynchronous concession, contrast similarity, conjunction	concession, contrast
Level 2, Threshold of 0.4		
Initial positively associated sense pairs	similarity, conjunction	None
Survived positively associated sense pairs	similarity, conjunction	None
No positively associated pairs were found for either dataset at level 2 with a threshold of 0.5		

5. Results Discussion

In this section, we will consider the four sense pairs that survived the pipeline for DiscoGeM leaf level at 0.3 as an example to validate the sanity of the decision pipeline results. Figure 18 shows the contingency tables for these four sense pairs.

Figure 18

Contingency Tables for the Four Survived Positively Associated Senses for DiscoGeM at Leaf Level with α = 0.3

precedence / synchronous	V	¬V
V	13	49
¬V	589	5854

arg2-as-denier / arg1-as-denier	V	¬V
V	23	98
¬V	330	6054

arg2-as-detail / succession	V	¬V
V	6	6
¬V	1410	5083

conjunction / similarity	V	¬V
V	60	20
¬V	2247	4178

Most instances fall into the 'no vote, no vote' category, where annotators consistently did not choose a particular sense.

The 'vote, vote' category indicates split decisions among annotators and has the fewest or nearly the fewest counts. For example, considering the conjunction & similarity pair, we can notice that out of 2,327 (60 + 20 + 2247) instances where there have been votes in favour of one or two of the sense pairs, there are:

- 60 instances where the annotators confused these two senses, and their votes were split.
- And 2,267 (2247 + 20) instances where annotators favoured one sense over the other.

This suggests that only 2.6% (60 / 2327) of the time, the annotators did not vote for the same sense, while 97.4% chose only one over the other.

The 'vote, vote' counts are particularly informative because they directly reflect instances of annotator confusion. To quantify the extent of the annotator confusion, a new indicator is proposed:

$$\text{Proposed indicator} = \frac{count(V,V)}{count(V,\neg V)+count\,(\neg V,V)+count(V,V)}$$

The proposed indicator's value spans from zero, indicating no confusion among annotators ('vote, vote' cell count is zero) to one, highlighting complete indecision. This metric captures the proportion of instances in which a significant number of annotators (more than six for $\alpha = 0.3$) were split in their decisions between sense pairs, thereby providing a measure to assess levels of confusion. Figures 19 and 20 show the proposed indicator values for the DiscoGeM dataset at the leaf level and level 2 with $\alpha = 0.3$.

We can notice that at the leaf level, as an example, the maximum value of the proposed indicator is only 10.8%. This offers a different interpretation of the annotator's behaviour that challenges previous findings.

Note that proposed indicator values for all 12 datasets can be found in the 3_results\binary\analysis_value_matrices\csv_files directory.

Figure 19

Proposed Indicator Values for the DiscoGeM Dataset at the leaf level with α = 0.3

	synchronous	precedence	succession	reason	result	arg1-as-goal	arg1-as-denier	arg2-as-denier	contrast	similarity	conjunction	disjunction	arg2-as-instance	arg1-as-detail	arg2-as-detail	arg2-as-subst
synchronous		2.0%	0.0%	0.0%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%	0.0%	0.3%	0.0%
precedence			0.0%	0.2%	6.4%	0.0%	0.1%	1.8%	0.3%	0.0%	4.4%	0.0%	0.3%	0.1%	0.6%	0.2%
succession				0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.3%	0.0%	0.4%	0.0%
reason					1.4%	0.0%	0.4%	0.4%	0.8%	0.0%	2.3%	0.0%	1.7%	0.8%	5.1%	0.2%
result						0.0%	0.8%	2.8%	0.8%	0.3%	10.8%	0.0%	1.2%	1.4%	4.4%	0.1%
arg1-as-goal							0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
arg1-as-denier								5.1%	1.7%	0.0%	0.3%	0.0%	0.0%	0.0%	0.8%	0.0%
arg2-as-denier									3.1%	0.0%	1.8%	0.0%	0.7%	0.5%	0.7%	0.0%
contrast										0.4%	1.0%	0.0%	0.2%	0.0%	0.6%	0.5%
similarity											2.6%	0.0%	0.7%	0.7%	0.2%	0.0%
conjunction												0.0%	2.2%	0.2%	10.7%	0.0%
disjunction													0.0%	0.0%	0.0%	0.0%
arg2-as-instance														0.4%	7.2%	0.0%
arg1-as-detail															0.7%	0.0%
arg2-as-detail																0.2%
arg2-as-subst																

Figure 20

Proposed Indicator Values for the DiscoGeM Dataset at the Level 2 with α = 0.3

	synchronous	asynchronous	cause	purpose	concession	contrast	similarity	conjunction	disjunction	instantiation	level-of-detail	substitution
synchronous		1.9%	0.5%	0.0%	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%	0.2%	0.0%
asynchronous			5.5%	0.0%	1.8%	0.2%	0.0%	5.0%	0.0%	0.5%	1.1%	0.2%
cause				0.0%	4.0%	0.9%	0.2%	13.1%	0.0%	2.6%	12.4%	0.2%
purpose					0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
concession						4.4%	0.1%	3.4%	0.0%	0.6%	2.5%	0.0%
contrast							0.4%	1.0%	0.0%	0.2%	0.6%	0.5%
similarity								2.6%	0.0%	0.6%	0.2%	0.0%
conjunction									0.0%	2.3%	12.0%	0.0%
disjunction										0.0%	0.0%	0.0%
instantiation											7.0%	0.0%
level-of-detail												0.2%
substitution												

6. Findings and Conclusion

1. The report introduces a decision pipeline developed to identify sense pairs that are positively associated with each other.
2. An indicator was proposed to check the sanity of the pipeline results concerning this report's objective.
3. The study found some sense pairs that show a statistically significant positive correlation.
4. However, these pairs are not consistent across different datasets and threshold values.
5. Moreover, most counts in the contingency tables belong to the cells where no one voted, most likely driving the positive correlation for these sense pairs.

In conclusion, the report recommends cautiously taking the results. This is because the cells with no votes, the 'not vote, no vote' cells, need to be disregarded and the cells that have votes, with particular attention to the ones that reflect confusion, the 'vote, vote' cells, need to be focused on.

However, The nature of the experiment drives the large counts for 'no vote, no vote' cells since only ten annotators each voted for one label out of 27 or 28 choices for each instance. Therefore, the positively associated sense pair resulting from this study is probably driven by this large count, which makes this association meaningless and does not reflect the report's objective. This is particularly evident from the results of the proposed indicator. Thus, the positively associated sense pair resulting from this study should be considered cautiously.

Appendices

Appendix A: Calculation Verification Sheet

In this Excel sheet in the manual_verification directory, you can select the corpus, the alpha threshold for binary transformation, and the sense pair you want to verify the calculation for. Note that Fisher's Exact Test is not calculated in this sheet. Moreover, it only accounts for the leaf level at present.

If any of the 'Expected Values' Table cells is less than 5, that cell is highlighted to indicate that the Chi2 and P-value (although calculated in this sheet) should not be considered. For this purpose, we have used Fisher's Exact Test. Although implemented in the code, it is not included in the value verification sheet.

Furthermore, you can browse through various sheets in the Excel file to see how each method was calculated. Please note that this sheet is independent of the code base and created manually. However, it uses the same calculation logic as the code. The values generated from this sheet are only for verification purposes.

	A	B	C	D
1	**update the grey cells**			
2	**Corpus**	discogem		
3	**Alpha threshold**	0.3		
4	**sense 1 (s1)**	synchronous		
5	**sense 2 (s2)**	precedence		
6				
7	Chi2	10.23	Do not use if any of the	
8	**Chi2 P value**	0.00	Expected Table cells is	
9	**OR**	2.71	highlighted	
10				
11	**YuleQ**	0.45		
12	**PMI**	1.18		
13				
14	**P(s1 given s2)**	0.02		
15	**P(s2 given s1)**	0.21		
16				
17				
18	**Contingency table**		s2	precedence
19			**¬V**	**V**
20	s1	**¬V**	5854	589
21	synchronous	**V**	49	13
22				
23	**Expected Table**		precedence	s2
24			**¬V**	**V**
25	synchronous	**¬V**	5847	596
26	s1	**V**	56	6
27				

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

Please note the following information about the data directory structure:

- The "0_raw_data" folder contains the original CSV files for the DiscoGeM and QADC datasets, which are two distinct datasets.

The "1_ready_to_transform" folder contains the cleaned and grouped data ready for transformation into binary format. This folder contains four datasets; more details can be found in section 3.a of the documentation.

The "2_ready_to_process" folder contains the data transformed into binary format. It contains 12 datasets; more details can be found in section 3.b of the documentation.

- The "3_results" folder contains multiple subfolders. The "Summary_report" subfolder contains the results for each statistical method used, which contains the results for the 12 datasets with positive and negative associations. The "Binary" subfolder contains three subfolders: "Contingency_tables" for the 12 datasets, "Expected_tables" for the 12 datasets, and "Analysis_value_matrices," which contains a value matrix for each statistical method for each dataset.