**Confusion Matrix and Inter-coder Agreement**

A screen shot of a social media post

Description automatically generatedA picture containing table

Description automatically generated

There were a total of 68 agreements and 32 disagreements out of the 100 sentences. Mathmatically, the Cohen’s Kappa Score is 0.5583. Statistically, the score may vary between **0.5782 to 0.7678** at 95% CI, suggesting a **moderate to substantial** inter-coder agreement.

**Major Types and Causes of Disagreement**

1. **The Lack of Context and Domain Knowledge:** Due to the lack of domain specific knowledge, the annotations could be slippery and ambiguous. For example, phrases like “risk factors”, “early tumor” can behave differently with varying contexts, affecting the annotations as they could sometimes be annotated as “Correlational” and “Causal” at others.
2. **Unclear Boundary:** The degree to which the two observed events have a causal relation, even though the statistically significant correlation exists, still remain unclear.
3. **Ambiguous Language:** The researchers are prone to seek for patterns conveying potential causality. Words like “cause” and “result” could be lucid. Words like “affect” and “link”, however, could be ambiguous in different scenarios.
4. **Varying Annotator Perspectives:** The inbuilt bias of the annotators was one of the main reasons for disagreements. Each annotator brought in different perceptions and even though there were ground rules for annotations, the abstract nature of the activity gave raises to a difference in annotations.

**Rationale for Resolving the Disagreement**

Certain ground rules were accepted for each of the categories. The disagreements were debated, the patterns were studied, and finally consensus was reached.

1. **No Relationship:** Sentences that contain just one entity, implying that there is no explicit cause-action/correlation pair.
2. **Correlational Relationship:** Sentence focus the degree or level of association between or the expected rate of appearance rather than the existence of cause-effect relation.
3. **Conditional Causal:** Sentence contains at least 2 entities with a clear cause-action pair and key words (e.g. might, may) that make the difference between direct- and conditional-causal.
4. **Direct Causal:** Sentence has the presences of a clear cause-action pair, including causative words or phrases that conveying cause-effect relation (e.g. result in, cause).

**Reproducibility and Limitation**

Even though inevitable error might occur during annotating, the schemas or language rules to determine causality and correlational relation is reproducible. For the recognition of causality, transitions (e.g. consequently, therefore), conjunctions (e.g. due to, since, because of), verb phrases (e.g. cause, result in, lead to) could be useful language hints. According to the language cues elaborated by Yu et. al. (2019), words or phrases as “association, predictor, at high risk of” could be used to identify correlational sentences. These rules can be replicated to any data set irrespective of the context with a fair accuracy. But, a clear understanding of the domain knowledge will definitely improve the accuracy of the annotations.

**Reference:**

Yu, B., Li, Y., & Wang, J. (2019, November). Detecting Causal Language Use in Science Findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4656-4666).