Inter annotator Agreement

About each entity, we have:

① Span (start index + end index)
↳ 212 - 248

② Value
↳ "THE SUPREME COURT OF BC"

③ Labels / Entity Name
↳ "COURT_NAME"

Possibilities of errors:

① A correct entity was not tagged at all

② The tagged entity was not an entity (According to the reviewer)

③ An entity was tagged but labelled incorrectly

④ A correct entity was tagged but partially

Case ①, ②, ③ are obvious → All are errors

In case ④,
if there is an exact match between the tagged entities, then only we assume it was tagged correctly, otherwise it was an error.

For tagged entities:
If at least two reviewers say that the entity was correctly tagged → We assume it was correctly tagged.

For missed entities:
If all the three reviewers tag a missed entity → We assume it is actually a correct entity but missed by the annotator.

(Ex.)

For Annotator 1

Annotator 1 → A1 ; Reviewer 1 → R1, R2, R3

		A1	R1	R2	R3	Agreement
Tagged by Annotator	TE1	✓	✓	✓	✓	4/4
	TE2	✓	✓	X	X	2/4
	TE3	✓	X	X	✓	2/4
	⋮	⋮				
	TE100	✓	✓	✓	X	3/4
Missed by Annotator	NTE1	X	X	✓	X	1/3
	NTE2	X	✓	✓	X	2/3
	NTE3	X				
	⋮	⋮				
	NTE10	X	✓	X	✓	2/3
Total		~~100~~ 10	~~90~~ 20	~~80~~ 30	~~70~~ 40	

Total tagged : 100

50 were tagged by 4 } 50 + 20 = 70
20 were tagged by 3 }

20 were tagged by 2 } 20 + 10 = 30
10 were tagged by 1 }

Total Missed : 10 (But tagged by Reviewers)

3 were tagged by 3 } 3
4 were tagged by 2 } 4 + 3 = 7
3 were tagged by 1 }

Total NEs in "Gold Standard" : 70 + 3 = 73
Total tagged by A1 : 100
Total correctly tagged by A1 : 70
Total Not tagged by A1 : 3

$$\text{Precision} = \frac{70}{100} \quad ; \quad \text{Recall} = \frac{70}{73}$$

F1 = whatever.

Similarly calculate for all annotators.