**AntMover Analysis**

**Corpus**: Elsevier 110 documents (OASTM)

Annotation Scheme Label

|  |  |
| --- | --- |
| Annotation Id | Annotation Scheme Name |
| 1 | Claiming centrality |
| 2 | Making topic generalization |
| 3 | Indicating a gap |
| 4 | Announcing present research |
| 5 | Announcing principal findings |
| 6 | Evaluation of research |

**1. Frequency of each Annotation Scheme**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Annotation Id | 1 | 2 | 3 | 4 | 5 | 6 |
| Total | 27 | 3725 | 96 | 2680 | 18186 | 6216 |



**Result**: Since there is a total of 110 document but there is only 27 of annotation id 1. It means there are some documents not being annotated with annotation id 1 (claiming centrality). The table below shows the number of documents not having specific annotation schemes.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Annotation Id | 1 | 2 | 3 | 4 | 5 | 6 |
| Exist | 22 | 109 | 51 | 109 | 110 | 110 |
| Not exist | 88 | 1 | 59 | 1 | 0 | 0 |



**Result**: Some documents are not annotated with certain annotation schemes. Human annotations should be done in order to verify whether these certain annotations should really exist/not exist in certain document. This is because the these all class of annotations, {1 to 6} should be found in each journals/paper as these annotation quite common.

**2. Frequency (proportion) of each Annotation Scheme**

I tried to find the frequency of each annotation scheme but in proportion to the number of annotations in a document id.

Formula:

(number of specific annotation in a document/ total annotation in a document)



Result: The frequency (proportion) chart is the same as the frequency chart. So, either method is the same.

**3. Is every sentence annotate by AntMover?**

By using the below SQL query, it shows that all sentences are annotated by AntMover.

*SELECT COUNT(\*) FROM SENTENCE WHERE sentence\_id*

*NOT IN(SELECT sentence\_id FROM SENTENCE\_ANNOTATION WHERE tool\_id=1);*

**Query result**: 0

**4. With the same corpus, is AntMover annotation process remain stable (reliable) for all the time?**

As suggested, I ran two time of annotations with the same corpus (110 articles) but in different session. Then, I tried to evaluate each session and found out that **each session produce the same result**. So, no statistical analysis require or else I will use **Krippendorf’s alpha** test for reliability

SQL query:

*#to count the number of sentences that having same annotation id for two different dates*

*SELECT COUNT(\*) FROM SENTENCE\_ANNOTATION sa1, SENTENCE\_ANNOTATION sa2*

*WHERE sa1.sentence\_id=sa2.sentence\_id*

*AND sa1.annotation\_id = sa2.annotation\_id*

*AND sa1.tool\_id = 1 AND sa2.tool\_id=1*

*AND sa1.sentence\_date='2017-10-08'*

*AND sa2.sentence\_date='2017-10-10';*

**Query Result:** 30930 (this is the total number of lines in 110 documents)

#to count number of sentences with different annotation id for two different dates

*SELECT COUNT(\*) FROM SENTENCE\_ANNOTATION sa1, SENTENCE\_ANNOTATION sa2*

*WHERE sa1.sentence\_id=sa2.sentence\_id*

*AND sa1.annotation\_id != sa2.annotation\_id*

*AND sa1.tool\_id = 1 AND sa2.tool\_id=1*

*AND sa1.sentence\_date='2017-10-08'*

*AND sa2.sentence\_date='2017-10-10';*

**Query Result:** 0

**5. AntMover sentence processing function (button: process) efficiency**

AntMover sentence processing function (a process that separate sentences in paragraph) is not very efficient as it cannot identified the end of sentences when the sentences end with following “strings”

a. <A character>.

Example: Figure A.

b. Abbreviation

Example: etc. or id.

**Result**: I need to do the text preprocessing before feed into AntMover software. Once the AntMover software able to identify each sentences in a document, it is able to annotate it but the correctness needs to be verify with annotations done by specialist (human) with the specific set of annotation scheme used by AntMover which is CAR model.