**Text Mining Short Textual Data for Ontology Term Mapping**

**Annotation Guidelines for Evaluation**

Gurinder Gosal

\*These Annotation Guidelines go with the evaluation sheet (named as *Annotator[No]\_EvaluationSheet*) distributed to different annotators. The evaluation sheet consists of two worksheets inside- one named *Annot[No]\_Distributed* and other named as *AccompanyingSheet\_Annot[No].*

First one is the actual sheet to be used for evaluation and the second one is only for administrative purposes and annotators should not use it (used for making dropdown lists only).

\*You can use website ***“https://www.ebi.ac.uk/ols/index”*** for looking terms from different ontologies used for evaluation exercise.

\*There are many cells which are non-editable in the evaluation sheet and they are shaded also in the Part-B of Evaluation Sheet. Which cells are these and the reasons for that are explained in these guidelines wherever required.

**PART-A of Evaluation Sheet**

All the system generated/prepared data is available in PART-A of the evaluation sheet through ***Column A to I*** and it is made non-editable so that the annotators are not able to change this data anyways.

**Columns A and B:** Entitled “*Annotation\_Id****”*** and “*Encryption\_Id*” are just Ids for different identification purposes and the annotators could ignore these.

**Column C:** Entitled “*Sample\_Desc****”*** describes the original sample description that is being mined and mapped with different ontology terms.

**Column D:** Entitled “*Cleaned\_Sample****”*** describes the cleaned sample description after preprocessing and other treatments.

**Column E:** Entitled “*RetainedSet\_MappedTerms\_with\_Resource\_IDs****”*** contains the mapped terms from sample descriptions along with ids of Resources (ontologies). This is used by the annotators to look at the mapping terms and their respective resources with ids.

**Column F:** Entitled “*Match\_Status (Macro Level)****”*** contains the match status at the high level. The mapping step in the pipeline results in three main types of matches at this level for input text to the resource terms.

* Full Term Match
* Component Match
* No Match

**\*(See Appendix A for further details)**

**Column G:** Entitled “*Match\_Status (Macro Level)****”*** explains the match status at the description level. The different processes, treatments and rules are reflected in this status description. This enables the better understanding and reasoning behind the matches, thereby facilitating the evaluation marking.

**\*(See Appendix B for further details)**

**Note: The column H and I are introduced after discussion with Will and were not part of the earlier presentation in the lab.**

**Column H:** Entitled “*Different Components (In case of Component Match)****”*** describes the different components assigned with component numbers along with ids of Resources (ontologies). As evident, this applies only to *“Component Match”* of “*Match\_Status (Macro Level)”.* This is introduced for and would be beneficial in identifying the partial matches and also pinpointing the errors to the corresponding components.

**Column I:** Entitled “*No. of Components (In case of Component Match)****”*** describes the number of components identified and matched in the sample descriptions. Again this applies only to *“Component Match”* of “*Match\_Status (Macro Level)”.* This would help in calculating the degree of partial matches after the evaluation.

**Column J:** This column is just serves as a divider between Part-A and Part-B.

**PART-B of Evaluation Sheet**

All the evaluation work has to be carried out in PART-B of the evaluation sheet through ***Column K to AA.***

**Column K:** Entitled “*(In cases of Full Term Match and No Match) Whether There is an Error in the Mapping (Y/N)****”*** gives the option of selecting **Yes or No** from a dropdown list in the cell to describe whether there is an error or not.

Please note, this applies only to *“Full Term Match or “No Match” cases* of “*Match\_Status (Macro Level)”.* Therefore, the cells in this column for the *“Component Match”* of “*Match\_Status (Macro Level)” are non-editable. (*Because the Component Match cases are dealt with in other columns due to involvement of multiple components – will be explained in the forthcoming discussion).

**Column L:** Entitled “(If Yes, What Kind of Error it is) [A,B,C,D or E]***”*** gives the option of selecting the type of error in the case of error identified by annotator being **Yes**. This option is again provided by the dropdown list available in the relevant cells of the column. Of course, in case of **No** in the previous column cell, the cell here should remain blank.

**\*The types of errors are described in the Appendix C.**

**Column M:** Entitled “In case of Errors B or E, Description of the error***”*** gives the annotator space for describing the error more in the cases of error types B and E.

Again note that for columns L and M, all above applies only to *“Full Term Match or “No Match” cases* of “*Match\_Status (Macro Level)”.* While, the cells in this column for the *“Component Match”* of “*Match\_Status (Macro Level)” would be non-editable.*

**Column N:** Entitled “Whether the New Terms and Processes, if any, are Valid? Y/N”gives the option of selecting **Yes or No** from a dropdown list in the cell to describe whether the new candidate terms for FOODON ontology or other ontologies (you would find them having **FOODON\_CandidateTerm** or **Other\_CandidateTerm** in resource Ids) are valid or not. Similarly, newly identified candidate processes (you would find them having [? **PROCESS]** in resource Ids) could be validated whether they are valid or not. This step actually helps in curating the new candidate terms/ processes which could be included in resources later on, once validated.

Please note, the cells in this column for the part other than new candidate terms/ processes would remain non-editable.

**Column O:** Entitled “In case of N, the reason for not being valid terms/processes”gives the option of describing the reason in more detail in the case of new terms/processes being identified as not valid.

Please note, the cells in this column also -for the part other than new candidate terms/ processes would remain non-editable.

**Column P:** Entitled “(In Case of Component Match) Whether the Match could be considered as Full Term Match Equivalent(Y/N)” gives the option of selecting **Yes or No** from a dropdown list in the cell to describe whether the given matches in the case of *“Component Match”* of “*Match\_Status (Macro Level)”* could be considered equivalent to Full Term Match.

* **Explanation-**There can be the cases when all the words of the input text match with different terms of resources as components and nothing is left without matching. Theses kind of matches are those component matches which could be **equivalent to the full term matches**.
  + For example, “Human(Feces)” sample description has its components matched to terms [homo sapiens:NCBITaxon\_9606, 'feces:UBERON\_0001988] and therefore could be considered as *Full Term Match* *Equivalent*.

Please note, the cells in this column for part describing *“Full Term Match or “No Match” cases* of “*Match\_Status (Macro Level)”.* would remain non-editable.

**Column Q:** Entitled “In case of N, the reason for not being Full Term Match Equivalent (e.g. Error in Component or Other)” gives the option of describing the reason in more detail in this case.

Please again note, the cells in this column part describing *“Full Term Match or “No Match” cases* of “*Match\_Status (Macro Level)”.* would remain non-editable.

**Column R to Z (9 columns)** give the possible 3 sets of annotations (optional) in the case of *“Component Match”* of “*Match\_Status (Macro Level)”.*

*One set (Column R, S and T) is explained and it should apply to the remaining sets also.*

**Column R & S:** Entitled “(In Case of Error in Component Match) Combination 1 (Error Type/Component No)” gives the option of annotating errors in the cases of component matches.

Here actually it is two part are combined together, *Error Type* and the *Component* to which error applies.

The annotator could use the dropdown list in the cell to describe the *error type [A, B, C, D or E]* in column **R** and use the dropdown list in the cell to describe the *component number [1..10]* (assuming there are no more than 10 components for a sample description match) in column **S**.

**Column T:** Entitled “In case of Errors B or E, Description of the error [ Combination 1 ]” gives the annotator space for describing the error more in the cases of error types B and E.

This similarly applies to two other sets, if needed.

**Note:** Assumption here is that at most the problems could be in 3 components (In case it is in more components, you can use the **Remark** column for further description)

Please note, the cells in this column part describing *“Full Term Match or “No Match” cases* of “*Match\_Status (Macro Level)”.* would remain non-editable for **columns R to Z**.

**Column AA:** Entitled “Additional Remark, if any” gives the annotator option to mention anything remarkable by writing in this column (e.g., the remark needed as mentioned above in the Note).

**APPENDICES**

|  |
| --- |
| **Appendix A:** |
| * **Macro Level (Match Status)**   The mapping step in the pipeline results in three main types of matches of input text to the resource terms.   * + Full Term Match   + Component Match   + No Match   **Full Term Match**   * This type of match results from the complete match of the whole chunk of text with some term in the resources.   + For example, a sample description, *crab meat* matches exactly and without any treatment with ontology term *crab meat:FOODON\_03311697*   **Component Match**   * This type of match results when there is not a complete match of the whole chunk of text with some term in the resources instead the different components of the text match with some terms in the resources. * There can be the cases when all the words of the input text match with different terms of resources as components and nothing is left without matching. Theses kind of matches are those component matches which could be **equivalent to the full term matches**.   + For example, “Human(Feces)” sample description has its components matched to terms [homo sapiens:NCBITaxon\_9606, 'feces:UBERON\_0001988] and therefore could be considered as *Full Term Match* *Equivalent*.   **No Match**   * This type of match results from no match of the input text with any term in the resources.   + For example, a sample description, ***hoy kom*** *does not* match with any ontology term |

|  |
| --- |
| **Appendix B:** |
| **Match\_Status (Micro Level)**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | **Sample\_Id** | **Sample\_Desc** | **Cleaned\_Sample** | **RetainedSet\_MappedTerms\_**  **with\_Resource\_IDs** | **Match\_Status (Macro Level)** | **Match\_Status (Micro Level)** | | samp3374 | peppermint tea | peppermint tea | [peppermint tea:FOODON\_03309568] | Full Term Match | {'A Direct Match'} | | samp3 | Chicken | chicken | [chicken:FOODON\_03411457] | Full Term Match | {'Change of Case in Input Data'} | | samp1408 | tuna, smoked | tuna smoked | [smoked tuna:FOODON\_03301592] | Full Term Match | {'Permutation of Tokens in Resource Term'} | | samp2173 | bovine meat | bovine meat | [bovine meat product:FOODON\_00001134] | Full Term Match | {'Change of Case of Resource and Suffix Treatment (Product) to Input'} | | samp2187 | dried shrimp | dried shrimp | [dried shrimp:(https://en.wikipedia.org/wiki/Dried\_shrimp)] | Full Term Match | {'Matching with Wikipedia Based Collocation Resource'} | | samp2394 | field cilantro | field cilantro | {'coriander:FOODON\_03411381', 'field:ENVO\_01000352'} | Component Match | {'Synonym Usage'} | | samp2472 | chili powder white | chili powder white | {'chili powder:FOODON\_03302030', 'white:[Quality-Color]'} | Component Match | {'Using Semantic Tagging Resources'} | | samp5 | cucumbers | cucumber | [cucumber:FOODON\_03411404] | Full Term Match | {'Inflection Treatment'} | | samp94 | porcine intestin; Sus scrofa | porcine intestine sus scrofa | {'intestine:UBERON\_0000160', 'porcine:zOther\_CandidateTerm\_155', 'sus scrofa:NCBITaxon\_9823'} | Component Match | {'Spelling Correction Treatment'} | | samp123 | frz shrimp | frozen shrimp | {'frozen:Process\_FOODON\_03470136', 'shrimp:FOODON\_03411237'} | Component Match | {'Abbreviation-Acronym Treatment'} | | samp357 | sambar powder | lentil curry powder | {'lentil:FOODON\_03411268', 'curry powder:FOODON\_03301842'} | Component Match | {'Non English Language Words Treatment'} | | samp2681 | spice/seasoning mix | spice seasoning mixture | {'seasoning:FOODON\_03316490', 'spice mixture:FOODON\_03304292'} | Component Match | {'Additional Match From POS Tagging Rule', 'Abbreviation-Acronym Treatment'} | | samp1576 | bovine minced beef | bovine minced beef | {'beef:FOODON\_03317335', 'minced:[MINCING PROCESS]', 'bovine:FOODON\_03414374'} | Component Match | {'Using Semantic Tagging Resources for Processes'} | | samp698 | 1 OPENED BAG - SELL BY 11/01/07 | 1 opened bag sell 11/01/07 | {'opened:[Quality]', '11/01/07:[DateOrDay]', 'bag:[Container-Or-Receptacle-Or-Enclosure]', 'sell:[Activity-Procedure]'} | Component Match | {'Using Semantic Tagging Resources', 'Using Semantic Tagging -[CARDINAL-ORDINAL]', 'Using Semantic Tagging -[DateOrDay]'} | | samp2158 | dairy farm lagoon stage 3 | dairy farm lagoon stage 3 | {'dairy:ENVO\_00003862', 'stage:[Quality]', '3:[CARDINAL-ORDINAL]', 'farm:ENVO\_00000078', 'lagoon:ENVO\_00000038'} | Component Match | {'Using Semantic Tagging Resources', 'Using Semantic Tagging -[CARDINAL-ORDINAL]'} | |

|  |
| --- |
| **Appendix C:** |
| * + **The Types of Errors to be Annotated**  |  |  | | --- | --- | | **Type ID** | **Type of Error** | | A | Wrong Term Mapping (semantically wrong) | | B | Error in Capturing by System | | C | Missed Mapping (Resource Term is Missing ) | | D | Error in Semantic Tagging | |  |  | |
| * + **Description of the Types of Errors to be Annotated**  |  |  |  |  | | --- | --- | --- | --- | | **Error Type** | **Sample\_Id** | **Sample\_Desc** | **Example (Hypothetical )** | | **Type A Error** | samp2198 | chile powder | Chile mapped to a country/GeoEntity | | **Type B Error** | samp3321 | frozen rabbits | For example, rabbits is not matched to 'rabbit:FOODON\_03411323' by the system | | **Type C Error** | samp1883 | hoy kom | hoy com {missed as there is no term in the resource] | | **Type D Error** | samp134 | corn snake cage swab | For example, cage maps to [Quality\_color] say instead of 'cage:[Container-Or-Receptacle-Or-Enclosure] | | **Type E Error** | samp1766 | COL050007 | For example, some description about this sample could be given | |

|  |
| --- |
| **Appendix D:** |
| **Evaluation (**Measures)   * + **Recall** – Recall measures the number of correctly identified items as a percentage of the total number of correct items. Recall is sometimes called as Coverage. The higher the recall rate, better the system is at not missing correct items.      * + **Precision** – Precision, sometimes also known as accuracy, measures the number of correctly identified items as a percentage of the number of items identified. The higher the precision, better the system is at ensuring that what is identified is correct.      * + **F-Measure** – F-Measure is the measure which takes into account both precision and recall and this combined one measure could be seen as the weighted average of precision and recall. F-Measure is also known as harmonic mean and usually in F-Measure, precision and recall are equally weighted, In that case F-Measure is known as F1 which means F-Measure set the value of weights equal to 1.     These above measuring terms can be calculated according to 3 different criteria – Strict, Lenient and average.   * + **Strict** – It considers all the partially correct responses as incorrect (or spurious/false-positives)   + **Lenient** – It considers all the partially correct responses as correct.   + **Average** – It allocates half weight of lenient to partially correct responses or can say it takes the average of strict and lenient. |

|  |
| --- |
| **Appendix E:** |
| **Semantic Tagging Lexicon**   * + Full Version ( Semantic Tags)     - [Abbreviation]     - [Cardinal-Ordinal]     - [GeoEntity]     - [Portion\_FoodOrOther]     - [Unit]     - [Quality] {[Quality-Color],[Quality-Food],[Quality-Shape],[Quality-Size],[Quality-State],[Quality-texture],[Quality-Time]..}   + [Activity-Procedure]   + [LocationContextual]   + [BodyPart-OR-OrganicPart]   + [Container-Or-Receptacle-Or-Enclosure]   + [DeadBody]   + [Equipment-OR-Device-OR-ManmadeObject]   + [Furniture]   + [GeoEntity]   + [GeographicArea-OR-Related]   + [Portion\_FoodOrOther]   + [Preposition-Containment]   + [Preposition-HavingOrigin]   + [Preposition-Presence]   + [Preposition-Support]   + [Structure-OR-Area]   + [Structure-OR-Area-OR-ManmadeObject]   + [Trademark]   + [WaterBody] |

|  |
| --- |
| **Appendix F:** |
| **Ontologies Covered as Resources**   |  | | --- | | **FOODON** | | **GENEPIO** | | **ENVO** | | **NCBITaxon** | | **UBERON** | | **UO** | | **PATO** | | **GO** | | **BFO** | | **BSPO** | | **CARO** | | **CHEBI** | | **ERO** | | **ExO** | | **FLU** | | **IAO** | | **MI** | | **OBI** | | **OGMS** | | **OMRSE** | | **PCO** | | **PO** | | **RO** | | **SYMP** | | **TRANS** | | **UBPROP** | |