PartML: Part-whole relation extraction

Version 0.1

Tuan Do N., Nick Moran, Mikel Mcdaniel

February 28, 2013

1 Annotation Guidelines: Explanation of PartML Tags and their Attributes

1.1 The tag <ENTITY>

An entity is any term which, in context, could posibly play the role of either "part" or "whole" or both. Therefore, any noun term, either physical (1) or abstract (2),(3) will be tagged, even if it is not involved in a part-whole relationship in the sentence. We exclude entities which cannot meet our criteria, such as action verbs and gerunds, geographical locations (but not political entities (3)), signal words (discussed later), etc.

1. The **octopus** inhabits many diverse **regions** of the **ocean**, including **coral reefs**, **pelagic waters**, and the **ocean floor**.

2. The **Catholic Church**, also known as the **Roman Catholic Church**, is the world's largest **Christian Church**, with 1.16 billion **members** worldwide. It is among the oldest **institutions** in **Christianity**, and has played a prominent **role** in the **history** of **Western civilisation**.

3. **OPEC** was founded to unify and coordinate **members' petroleum policies**. Between 1960 and 1975, the **organization** expanded to include **Qatar (1961)**, **Indonesia (1962)**, **Libya (1962)**, **the United Arab Emirates (1967)**, **Algeria (1969)**, and **Nigeria (1971)**.

1.1.1 How to annotate ENTITYs

We should employ the following strategies in annotating ENTITYS:

- In a nonimal group that expresses an entity, the ENTITY tag will not only apply to the head, but might include modifiers. Specifically, when these modifiers significantly add meanings to the head, to the extent that we could say the group of words denote an independent entity in the real world. For example:

  1. Over the approximately 900 years that **castles** were built they took on a great many **forms** with many different **features**, although some, such as **curtain walls** and **arrowslits**, were commonplace.
  In this instance, we apply the tag for the whole phrase "curtain walls" as it refers to a particular type of wall, and is considered as an independent entity.

  2. **Bigfoot** is described in **reports** as a large hairy **ape-like creature**, in a range of 6–10 feet (2–3 m) tall, weighing in excess of 500 pounds (230 kg), and covered in dark brown or dark reddish **hair**.

In this instance, we apply the tag for only "ape-like creature" instead of the whole noun phrase "large hairy ape-like creature", as we can see that adjective modifiers add some characteristics to the entity, but do not create a new and independent entity. On the contrary, modifier "ape-like" specifies the type of "creatures" in the context, and effectively makes "ape-like creatures" a particular entity. In other words, we could remove the determiner "a" and adjective "large" and "hairy" and the sentence still makes sense. However, if we remove "ape-like", we have lost key information about the entity.

- If the nominal is an apposition (which are constructed by two noun-phrase elements being placed side by side, with one element serving to define or modify the other), firstly check if the whole nominal could be considered an entity (like 'curtain wall' of (1) of the first rule). If it is, tag it as a single entity, otherwise tag it as two separate entities.

    1. *The ONLY **way** to accomplish **this** is to charge and try **President Barrack Obama**, **Defense Secretary Leon Panetta** and **anyone** else*

- The tags could also be applied to most pronouns, but not some special types of pronouns i.e. reflexive pronouns ("himself", "herself"), interrogative pronouns ("who", "what" in interrogative sentences). We assume that samples in the corpus make clear the referent of any anaphora.

    1. *The last **one** has some **hostesses** but also have an **ice bar** and a live all thai girl **band** which is not too bad.*

    2. *In fact, **they** did so well that the lowest **score** for **Camel's candidates** in any **part** of the **exams**, **which** consisted of **theory**, **knowledge presentations**, was 4.8 out of 5*

- The tag should NOT be applied to a location mention if the mention is a proper noun referring to a unique location (both real (1) and fictional (2)). Nominals playing functional roles or to show direction such as *'middle'*, *'south'* etc. (2)(4) are also excluded from being tagged. However, some terms could be a location term in one context and an *object* term in another context, such as political entities or social relationship.

    1. *Paris is the **capital** and largest **city** of France.*

    2. *... to the **elf haven** of Rivendell, across the Misty Mountains and the **black forest** of Mirkwood, to Lake-town in the middle of Long Lake, and eventually to the Mountain itself*

    3. ***Oceanographers** divide the **ocean** into different **zones** depending on the present physical and biological **conditions**.*

    4. *... extends from the Arctic in the north to the Southern Ocean (or, depending on **definition**, to Antarctica) in the south*

- The tag should NOT be applied to a temporal term.

    1. *... the **Bolivarian movement** from the beginning had a strong **class component** ...*

    2. *In the **Gregorian calendar**, New Year's Eve (also Saint Silvester's Day in many **countries**), the last day of the year, is on December 31.*

- The tag should NOT be applied to quantity or measurement terms. For example:

    1. ***Absinthe** has a high level of **alcohol***

- The tag should NOT be applied to gerunds (a form of verb ending in *-ing* functions as a verbal noun). Pure nouns ending in *-ing* such as 'building', or 'reading' are not in this category.

    1. *Eating **biscuits** in front of the **television** is one **way** to relax*

    2. *... the **traffic** alone will tell **you** that Christmas shopping is rapidly reaching **its peak**.*

1.1.2 BNF for the ENTITY tags

attributes ::= eid type reference_count [comment]
eid ::= e<integer>
type ::= 'CONCRETE' |'ABSTRACT' |'OTHER'
reference_count ::= 'SINGLE' |'MULTIPLE' |'MASS' |'ZERO' |'AMBIGUOUS'
comment ::= CDATA

1.1.3 Attributes for ENTITYs

a. **Entity ID number** (eid) Non-optional attribute. Each entity has to be identified by a unique ID number. This will be automatically assigned by the annotation tool every time an ENTITY tag is assigned to some string.

b. **Type**: Non-optional attribute. Each enity belongs to one of the following classes.

- CONCRETE: Entities of concrete class refers to physical objects, that could be observed by at least some sense.
- ABSTRACT: Entities of abstract class refer to abstract objects; that is, ideas or concepts.
- OTHER: Entities that are not easily classified into two classes above.

c. **Reference count**: Non-optional attribute. Each event belongs to one of the following classes:

- SINGLE: This class applies to a nominals which are singular in the given context.
- MULTIPLE: This class apply to nominals which are plural in the given context.
  It should be noted that some nominals are morphologically plural, but should be tagged as singular. Those words include words of foreign origins, such as 'spaghetti', 'paparazi', 'algae' etc. (1)
  Vice versa, some are morphologically singular, but should be tagged as multiple, especially when it is used grammatically as plural noun.(2) For example:

  (a) *will hound you like the <entity reference_count = singular>**paparazi**</entity >on bald headed Britney Spears..*
  (b) *... the <entity reference_count = multiple >**government**</entity >are worried our internet may overload, slowing browser speeds...*
- MASS: This class applies to uncountable nouns, mass nouns, abstract nouns such as 'water', 'sand' or 'money'.
- ZERO: This class apply to nominals that are explicitly non-existent or numerically zero. It will often apply to nominals with modifiers such as 'no', 'none' or 'zero'.
- OTHER: For entities that are not easily classified into the four classes above.

1.2 The <PARTOF>Link

A PARTOF link is a relationship to capture the meronymy relationship between a pair of ENTITYs. A PARTOF estabilshes a link between two entities if they are:

1. **Establish a part - whole or component - integral relationship**
   The 'whole' in a relationship could be divided into a number of 'part's, where each plays a structural or functional role to each other and to the 'whole'. In this kind of relationship, the 'whole' is always arranged in a patterned structure.
   This kind of relation is not only limited to physical entities (a), it could also be applied to abstract

entities and organizations if the relations between entities of these types also display a patterned structure of the 'whole' and functional or structural role of the 'part' (b).
For example:

<div align="center">part-whole</div>

(a) Dry **air**   contains roughly (by volume) 78.09%   **nitrogen**   ...

<div align="center">part-whole</div>

(b) In April 1951,   **Allied Command Europe**   is established as an integrated military structure for   **NATO**   .

2. **Establish a member - collection relationship**
   The difference between member - collection relationship and part - whole relationship is that 'member' entities are not required to have structural or funtional role of the 'collection'. The 'collection' is normally a group of homogenous 'member's that are grouped together because of their spatial proximity or social connections.

<div align="center">member-collection</div>

   (a) **SNG** is a regional **organization**   **whose**   participating   **countries**   are former Soviet **Republics.**

3. **Establish a subtance - object relationship**
   In this relation, the 'subtance' is an inseparable portion of the 'object'. The 'subtance' is normally a material or chemical subtance that the 'object' contains. Inseparibility in this sense only regards the current context, it doesn't mean that another object that have the same generic term as this 'object' or a similar object must also compose of this 'subtance'. For example:

<div align="center">subtance-object</div>

   (a) The **bike**   is partly   **steel**   .

1.2.1   How to annotate PARTOFs

We should employ the following strategies in annotating PARTOFs:

- Be aware of the difference between meronymy relationships and spatial association. Some meronymy relationships also connote a sense of spatial association, like when 'engine' is part of 'a car'. However, a spatial association alone does not imply any connection between two entites other than the spatial one.
 The two most frequent types of mere spatial association are **inclusion** and **attachment**. Avoid tagging them with a PARTOF link.

  1. ... **your classmates** and **teachers** are in the **building** with **you** ...
  2. **Earrings** are attached to **ears**

- Be aware of the difference between meronymy relationship and possesion. For example:

  1. **I** has a **car**

- Be aware of the difference between the part - whole relationship and subtance - object relationship. There is a subtle difference between those classes. Ask yourself a question: *If X doesn't have Y anymore, is X still X?*. If yes, tag the link as 'part - whole', else tag it 'subtance - object'. For example:
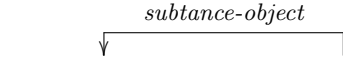
1. *subtance-object*

   Tequilla *contains* **alcohol** .

   Here if Tequilla doesn't have alcohol anymore, it would be some kind of juice, so the link should be 'subtance - object'.
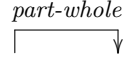
2. *part-whole*

   We have **tomato salad** .

   Here, if our salad doesn't have tomato anymore, it is still a salad, just with a different ingredients, so the link should be 'part-whole'.
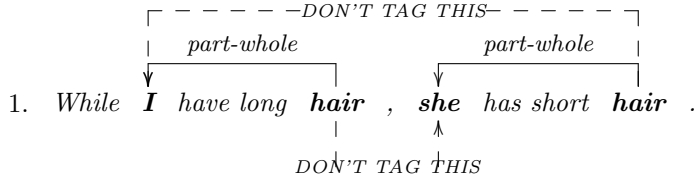
- Be aware of the difference between our meronymic relationship and portion - mass relationship (that we don't want to include in the scope of this work). The difference is that you can divide the mass into a theoretically infinte number of homogenous portions that still keep some characteristics of the mass. An example of that portion mass relation is:

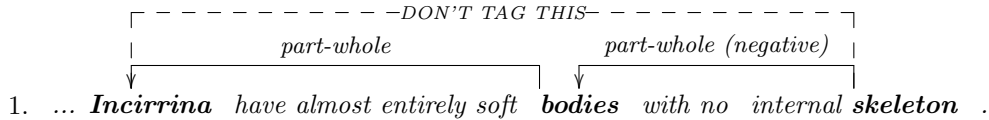   1. **They** *got a* <u>water</u> **sample** *from the* **lake**

- Be aware of the difference between member - collection relationships and member - class relationships which is not a meronym relation. (Intuitively, X is a class of Y if Y is an X. ex: A cat is a feline.) Example of member - class relationship is:

   1. *The* **car** *is a* **Ferrari**

- Exclusion of negative meronymic relationships inferred from a sentence of two contrastive parts. We consider that type of meronymic relationship to be indirect and ambiguous. Sentences of this type employ conjuntive words such as 'while', 'whereas', 'although'. For example:

   1. *While* **I** *have long* **hair** , **she** *has short* **hair** .
   (─ ─ ─ ─ ─DON'T TAG THIS─ ─ ─ ─ ─; *part-whole*; *part-whole*; DON'T TAG THIS)

- Exclusion of transitive meronymic relationship. When we could draw directly from the context that A is part of B and B is part of C, if there is a meronymic relation between A and C, do NOT tag it, unless the text explicitly indicates the relation between A and C (as opposed to merely implying it via transitivity). For example:

   1. *...* **Incirrina** *have almost entirely soft* **bodies** *with no* *internal* **skeleton** .
   (─ ─ ─ ─ ─ ─ ─ ─DON'T TAG THIS─ ─ ─ ─ ─ ─ ─ ─; *part-whole*; *part-whole (negative)*)

- There is a PARTOF link between two ENTITYs X and Y if the syntatically expressed meronymic relation directly involves those ENTITYs. If there is a coreferent mention of X, X', a PARTOF link between X' and Y should not be tagged, even though such a relationship could be inferred. For example:

   1. **John** *is tall.* **His** **hair** *is long.*
   (─DON'T TAG THIS─; *part-whole*)

5
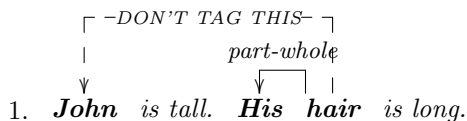
- Do NOT tag the link if the meronym relationship is not clear from the syntax and can only be implied by prior world knowledge. For example:

  1. *The higher the protein content the harder and stronger the* **flour**, *and the more it will produce crusty or chewy* **breads**.
     We know from prior knowledge that flour is a part of bread. However, it's not really clear from the sentence. A similar sentence about different ovens producing different types of bread would have the same structure, but ovens are not part of bread. Therefore do NOT tag the relation between flour and bread. (Intuitively, if you could substitute all of the entities with "X1", "X2", etc. and you can't tell if "X2" is a part of "X1", do not tag it.)

1.2.2   BNF for the PARTOF tags

attributes ::= pid relationship part_id whole_id negativity [signal_list] [comment]
pid ::= p<integer>
relationship ::= 'PART_OF' |'MEMBER_OF' |'SUBTANCE_OF'
part_id ::= IDREF
{part_id ::= eid}
whole_id ::= IDREF
{whole_id ::= eid}
negativity ::= 'POSITIVE' |'NEGATIVE'
signal_list ::= CDATA
{signal_list ::= sid(' 'sid)+}
comment ::= CDATA

1.2.3   Attributes for PARTOFs

a. **PARTOF ID number** (Pid) Non-optional attribute. Each PARTOF link has to be identified by a unique ID number. This will be automatically assigned by the annotation tool every time an PARTOF link is created.

b. **Relationship** Non-optional attribute. Assign value as above discussion.

c. **ID of 'part' entity** (part_id) Non-optional attribute. Value is entity ID of the 'part', 'member' or 'subtance' in the relationship.

d. **ID of 'whole' entity** (whole_id) Non-optional attribute. Value is entity ID of the 'whole', 'collection' or 'object' in the relationship.

e. **Negativity of relationship** (negativity) Non-optional attribute. 'POSITIVE' if the relations is affirmative, else 'NEGATIVE'.

f. **List of IDs of signal words** (signal_list) Optional attribute. Value is a list of signal word IDs separated by a single space character.

1.3   The tag <SIGNAL>

SIGNAL tag is applied to all words in the text that semantically signify the part - whole relationship. A part - whole relationship could have zero (1), one (2), (4), (5) or multiple (3) SIGNAL words.
SIGNAL words play an important role for any further work on meronymic relationship recognition.
SIGNAL tag could be applied to multiple types of words, including noun, verb (2), (3), adjective, past(4) or present(5) participle, prepositions (3) etc.

1. *part-whole*

   John *is tall.* **His hair** *is long.*

2. *subtance-object*

   Tequilla <u>contains</u>$_{SIGNAL}$ **alcohol** .

3. *part-whole*

   Oceanographers <u>divide</u>$_{SIGNAL}$ *the* **ocean** <u>into</u>$_{SIGNAL}$ *different* **zones** ...

4. *part-whole*

   ... **Allied Command Europe** *is established as an* <u>integrated</u>$_{SIGNAL}$ *military structure for* **NATO** .

5. *member-collection*

   ... **whose** <u>participating</u>$_{SIGNAL}$ **countries** *are* **former Soviet Republics.**

1.3.1 How to annotate SIGNALs

- If the SIGNAL is a phrasal verb, marked both the main verb and its particle or preposition.
  If the verb is not phrasal verb but the main verb and its preposition both signify the relation, they will all be marked.

1. *part-whole*

   Oceanographers <u>divide</u>$_{SIGNAL}$ *the* **ocean** <u>into</u>$_{SIGNAL}$ *different* **zones** ...

1.3.2 BNF for the SIGNAL tags

attributes ::= sid [comment] sid ::= s<integer>

1.3.3 Attributes for SIGNALs

a. **PARTOF ID number** (Pid) Non-optional attribute. Each PARTOF link has to be identified by a unique ID number. This will be automatically assigned by the annotation tool every time an PARTOF link is created.

References

Roser Sauri, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky, <u>TimeML Annotation Guidelines, Version 1.2.1</u>. 2006.

Morton E. Winston, Roger Chaffin, Douglas Herrmann, <u>A Taxonomy of Part Whole Relation</u>. Cognitive Science, Volume 11, Issue 4, pages 417–444, October 1987