PartML: Final Paper

Tuan Do N. Nick Moran Mikel Mcdaniel

May 11, 2013

Abstract

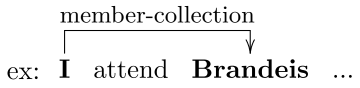
Detecting meronymic relationships is of great use in many natural language processing domains, such as processing biomedical text, question-answering systems, text summarization, and text understanding. (8) (10) We describe, PartML, an annotation specification and guideline for annotating part-whole relationships and related entity and relationship attributes that are useful for applying machine learning. We also provide a brief overview of previous works and their approaches for defining and automatically finding part-whole relationships in text, then discuss the strenghts and weaknesses of our annotation schemes, including inter-annotator agreement scores, and finally end with results of machine learning methods applied to our gold-standard corpus.

1 Introduction

PartML is an annotation scheme designed to capture meronymic, or part-whole, relationships and related attributes that are useful for applying machine learning to automatically detect these relationships in unannotated texts. To cover a large knowledge domain that is both wide and deep, PartML's guidelines are tuned to work well with annotated plain-text versions of English Wikipedia articles, but should be directly applicable or easily translatable to other languages and types of text.

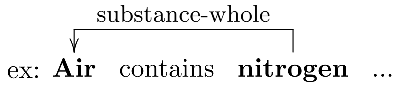
While different sources have varying views on the classes of part-whole relationships, we use the three basic types defined by WordNet:

1. part of (hand *and* arm)
2. member of (senator *and* senate)


member-collection

ex: **I** attend **Brandeis** ...

3. substance of (gold *and* ring)


substance-whole

ex: **Air** contains **nitrogen** ...

Our specification captures each of these three types with the goal of differentiating the exact type of meronym when one is detected. Our findings include differences in the agreement of annotation for these different types, as well as preliminary results from machine learning efforts. We describe potential improvements to our specification and guideline to improve agreement, to bolster future machine learning, and to expand the scope of the task to related areas.

2 Previous works

Part–whole or meronymy relations are an important binary semantic relation that have been worked on from the time of the atomists (Plato, Aristotle, and the Scholastics) and further investigated in the beginning of the 20th century. Though it is always considered a fundamental relation, there is no consensus on formal definition, representation, and classification and of meronymy relations in previous works. Studies of logic and philosophy of meronymy relations also disagree on its basic properties, especially the transitivity property (whether the relation is transitive or not).

In regards to formal representation and formulization of meronymy relations, there is a simple approach that models the relations as a transitive binary relation in Description Logic. That includes works of Artale et al. (1996) (1), Sattler 1995(2). For example, to say that car has wheels and that in turn the wheels have tires as their parts, it could represented as:

$$Car = \exists \succeq (Wheel \sqcap \succeq Tires)$$

lead to

$$Car \ \sqsubset \succeq Tires$$

Priss, in his discussion on construction of WordNet (3), used a formal definition of meronymy relations, together with other semantic relations (i.e. hypernymy, synonymy) based on Relational Concept Analysis. The relation is defined among synsets, that is differentiated from lexical relation (i.e. antonymy). His work characterizes the relation with three attributes irreflexive, antisymmetric and acyclic. The relation is parameterized by quantificational tags, that are simple quantifiers, such as *for all*, *exactly 1*, and *some*.

$$c_1 R^r[Q^1, Q^2;]c_2 :\iff Q^1_{g1 \in Ext(c1)} Q^2_{g2 \in Ext(c2)} : g_1 r g_2$$

$$c_1 R^r[;Q^3, Q^4]c_2 :\iff Q^3_{g2 \in Ext(c2)} Q^4_{g1 \in Ext(c1)} : g_1 r g_2$$

$$c_1 R^r[Q^1, Q^2; Q^3, Q^4]c_2 :\iff c_1 R^r[Q^1, Q^2;]c_2$$

$$and c_1 R^r[;Q^3, Q^4]c_2$$

There is also no agreement on the classification of part-whole relations. Cruse (4), using the same notion as Priss, described four subclasses, considering the cardinality of two word concepts in the relation. Iris et al.(5) specified four other classes: functional component (exactly one whole), Segmented whole/mass nouns/is-substance-of (multiple homogenous parts - one whole), membership relation (multiple whole), individual concepts: (one part - one whole). Although this approach captures the cardinality of involved objects in the extent of the concepts, which is fundamental in part-whole relationships, it ignores other important attributes such as the physical existance of these objects (discrete or abstract) that could also provide insight on the differentiation of meronymy relation. It also doesn't explicitly describe the inherently dependent or functional relation between involved objects in each subclass.

Winston et al. (1987) (6) described six types of meronymic relations that are (1) COMPONENT – INTEGRAL,(2) MEMBER – COLLECTION,(3) PORTION – MASS,(4) STUFF – OBJECT,(5) FEATURE – ACTIVITY, and (6) PLACE – AREA. In addition, they proposed three attributes of relation called *relation elements* i.e. functional, homeomerous, separable. For example, PORTION–MASS is homeomerous because portions are similar to each other, and separable as portion could be disconnected from mass (cut a slice from a pie). They also suggest a hierachical categorization of semantic relations, giving a distinction between meronymic relation and some other *'part-of'* relations, including spatial inclusion, class member, attributes and possesion.

Keet et al. (2007) distinguishs the use of two terms *part-whole relation* and *meronymic relation.* They include meronymic relation as a subclass of part-whole relation while the other class is a *mereological part-of relation.* The latter is a transitive relation that encompasses a wide range of functional and spatial part-of relations.

Most of previous works apply a predefined set of lexical-syntatic patterns, such as the following pattern from (8) to construct the meronymic corpus:

Cluster	Patterns	Freq.	Coverage	Examples
C1. genitives	NP_X of NP_Y			eyes of the baby
and	NP_Y's NP_X	282	52.71%	girl's mouth
verb *to have*	NP_Y have NP_X			The table has four legs.
C2. noun	NP_{XY}	86	16.07%	door knob
compounds	NP_{YX}			turkey pie
C3. preposition	$NP_Y PP_X$	133	24.86%	A bird without wings cannot fly
	$NP_X PP_Y$			A room in the house.
C4. other	others	34	6.36%	The Supreme Court is a branch of the Government.

This rule-based approach for extraction of meronymy relation suffers from insufficiency because many patterns that signify part-whole relations are complicated and involve longer structures.

3 Task

While many previous works focus on identifying meronymic relationships via rule-based and pattern-matching methods which have been boot-strapped from known meronymic pairs, our goal was to use subtler syntactic clues to the existence of these relationships with the hope of powering a learner which is able to handle a wider variety of structures indicating the sought after relationships. To this end, we frame meronym detection as a classification problem in which pairs of entities either are or are not in a meronymic relationship. It is important to note that this relationship is inherently directional, because we need to distinguish the part from the whole.

Although our initial machine learning efforts focus on simple detection of meronymic relationships, our specification is set up to also provide features for further classification of a detected relationship as one of the three types mentioned earlier. As such, the distinction between the three is important to our task, and its effective communication to annotators in the guideline is critical.

The first type of meronym is the simple part-whole relationship. In this sort of relationship, the entity playing the part role is a distinct component of the whole, in which it plays some structural role. While this is a very common type of meronym for physical objects, especially those with mechanical or anatomical structure, it can also occur in abstract entities, in which there is some sort of underlying pattern structure for the part to participate in.

The second type of meronym is the substance relationship. This is differentatied from the part-whole type by the fact that the entity in the part role is often homogenous and is the material out of which the part is made, rather than playing a specific structural role. It is possible for an entity to have be made of multiple substances, as is often the case for mixtures and conglomerates. Mass nouns are common in the part role for these substance-based relationships.

The third type is the member-group relationship. This type is similar to the part-whole relationship, but member entities do not have a direct structural role in the whole. Instead, this type of relationship is defined by the presence of an associative relationship, often based on physical proximity or social connections.

In order to keep a focused and well-defined task, we exclude certain relationships that other authors have included as meronyms, or which are tangentially related to meronyms. This include relationships of classification (often termed is-a) relationships, the containment relationship (a form of has-a), geographic and location-based relationships, mass-portion relationships, activity-based meronyms (we focus on nouns as entities, while these are based on verbs), possession and owernship relationships, and others. These are left open as possible extensions of the specification in future work.

As an aid to machine learning, we include explicitly negative meronymic relationships in our specification. These often show very similar syntactic structure to positive instances, except for the presence of a negation word. If these were simply excluded, they would present confusing instances for training.

Because our learning effort is focused on syntactic clues, we annotate only those relationships which are made clear from the text. We therefore do not annotate pairs of entities which world-knowledge tells use form a meronymic relationship, but which appear unrelated in the text. We also exclude those relationships which are not explicitly indicated, but could be derived by transitivity.

To further elucidate syntactic clues, annotators are asked to note signal terms, which they feel indicate the presence of the meronymic relationship. Common signal words and phrases include "made of", "part of", "contains", etc. These tagged signals can then be used as powerful features for machine learning.

Finally, we include a few simple attributes for our entities, which, while not directly related to the target phenomenon, are hoped to be useful as additional features for machine learning. We focus on the type of object described by the entity as well as its count (that is, whether it is singular or plural) as potentially beneficial, but future work could extend this to include many other attributes.

4 Specification

The core extent tag of our specification is the ENTITY tag. This tag is applied to nouns which participate in a meronymic relationship, and is used for both the meronym and the holonym in the relationshp. Only those nouns for which a meronymic relationship is syntactically indicated by the text are tagged – pairs which form a part-whole relationship, but for which that relationship is not specifically indicated by the text (that is, pairs for which the only indication of the relationship is prior world knowledge) are excluded.

Example: Tunicate [**larvae**$_{ENTITY}$] have both a [**notochord**$_{ENTITY}$] and a nerve [**cord**$_{ENTITY}$] which are lost in adulthood.

Each ENTITY tag has two attributes, type and reference_count. These attributes are not directly related to the phenomenon of meronymy, but are included as potential features for machine learning. We suspect that the nature of the entities involved in a meronymic relationship is relevant to determining the nature of that relationship. To that end, we annotate whether the entity is physical or abstract (covered by the type attribute), and whether the entity is singular, plural, a mass noun, or has zero count (covered by the reference_count attribute). The zero value is to be used in instances where a negative relation is expressed. For example, in the sentence from our corpus: "Speedway motorcycles use only one gear and have no brakes...", we would tag 'brakes' as having zero count.

Our other extent tag does not participate directly in the meronymic relationships, but captures SIGNAL words which the annotator feels give syntactic clues about the presence of such a relationship. Certain terms or phrases, such as "contains", "made of", "consists of", and the obvious "part of", often indicate that one entity is part of another. The goal of this extent tag is to capture these indicative terms for use as features during machine learning. Because they are not directly involved in our target phenomenon, we did not include any additional attributes for this tag.

Example: The bat is [**made of**$_{SIGNAL}$] wood.

In addition to our extent tags, we use a link tag, PARTOF, to capture which pairs of entities are participating in a meronymic relationship, and what the directionality of that relationship is. Each link joins two ENTITY tags, as well as any number of SIGNAL tags which are relevant to that link. One of the two linked ENTITY tags is identified as the 'to' member of the link, which is the object that constitutes the 'whole', and the other is the 'from' member, which is, of course, the 'part'.

Example:

PART_OF

Carousels are populated with **horses.**

Further, we include an attribute on the PARTOF tag which indicates which handles our secondary task beyond identification of meronyms - the classification of them. The relationship type attribute codifies the link as being one of our three types (part-whole, substance, and member). We also include a rarely-used 'other' option as a convenience for annotators to indicate uncertainty. However, the intent would then be to resolve this issue, rather than leave an 'other' link in the corpus for learning.

5 Corpus

5.1 Collection

Our corpus was collected from a variety of English Wikipedia articles, stripped of formatting and non-textual elements. Wikipedia was chosen because it provides several key features that we felt were important in a corpus. First, Wikipedia articles provide coverage of many domains, and detailed information is available for each domain. Additionally, Wikipedia's articles are written in a consistent, formal style which features an indicative voice and gramatically sound complete sentences. This provides clear and accurate text for our annotators, and consistent features for machine learning.

Gathering of the corpus was split between two approaches. The first was automatic collection loosely modeling that seen in prior works, in which articles were searched for a sufficient density of known signal terms. While this provided a number of useful articles, it was also prone to generating false-positives that did not have any instances of our target phenomenon.

The second method was human-driven, and used Wikipedia's built-in random article feature to generate a selection of candidates. Although many had to be discarded for being extremely short, of low quality, or lacking meronyms, we were able to salvage candidates by navigating to a related but more general article. These articles were often more detailed and of higher quality as they were more frequently used and editted. For example, although an article about a specific type of fly might be a mere stub containing no useful text, the related article about insects would have a wealth of useful information.

5.2 Statistics

Our corpus (counting only those documents which were annotated) contains 40 documents, with an average of roughly 425 words per document. Document length varies significantly, from only 71 words for an article about the Tarbell Building to nearly 1,000 words for an article about Carousels. Some articles are quite dense with entities and links, with one article containing nearly 70 links, while others are extremely sparse, or, in a few cases, entirely devoid of links. The frequencies of various numbers of links is shown below:

Links	Count
0-10	20
11-20	14
21-30	3
31-40	2
> 40	1

One of the reasons for this wide variety of densities is the need to find articles representing all three types of meronyms. Articles which deal with mechanical or anatomical parts often have long lists of components, because they aim to explain the parts of the subject. This results in many entities and many links. By contrast, subjects that are more abstract rarely have this sort of detailed breakdown of their components, and so entities and links are rare in those articles.

As a result of the tendency for different types of meronym to show up with different frequencies, the number of each type in our corpus is not balanced. Part-of relations are much more common, while substance-based meronyms are quite rare. Often, items have a single homogenous substance (as in a ring made of gold, for example), while they might have many mechanical parts. The relative frequencies of the different types identified by annotators are shown below (the frequencies in the adjudicated gold standard are quite similar):

Type	Percentage
Part-whole	63%
Member	27%
Substance	9%
Other/uncategorized	<1%

Because articles that have strong examples of non-physical meronymic relationships are comparatively rare, and comparatively sparser, our examples of these were generally of lower quality. It was easy to find a wealth of good examples for physical meronyms, but for many abstract topics we had to settle for samples that were perhaps more ambiguous and prone to annotator disagreement. This does bring some benefit, however, as these difficult examples provide a richer training set for future machine learning.

6 MAMA

Our original specification differed from our current one in two important ways. First, instead of annotating only those entities which actually participate in a meronymic relationship, we annotated all nouns in the text which were eligible to participate in such a relationship, even if they did not. This resulting in marking nearly every noun as an entity, with the only ones excluded being those that could only participate in a type of meronymic relationship which we were not focusing on, such as gerunds (which would be in activity-based relationships) or locations (which would be in geographic-based relationships). The goal of this was to ensure that we had an accurate representation of not only the positive examples of our phenomenon, but also the negative examples.

However, we found that the number of non-participating nouns in the text vastly outweighed the number of actual meronyms, and so annotators spent almost all of their time marking negative examples. Although these negative instances may be somewhat helpful for machine learning, we felt it was more important to get good positive examples, so the annotators' time would be better spent on those. As such, we revised our specification to its current form in which only those entities actually in a meronymic relationship should be tagged. By simplifying the task in this way, we allowed annotators to generate a much greater volume of useful data.

Ultimately, we were still able to generate a sufficient number of negative examples by taking pairs of tagged entities which participated in meronymic relationships, but not with each other. Because the number of possible pairings from a set is quite large, this resulted in a very large number of potential negative examples. Further negative examples could be generated by applying a part-of-speech tagger to the corpus and extracting nouns which were not tagged as entities. This second method may be more useful in the long-run, as it would mirror the process required to apply a classifier to completely untagged text.

The second important change to our specification was the way we dealt with noun phrases as entities. We originally felt that, in some cases, modifiers of the head of the phrase significantly alter its meaning, such that it would be necessary to have those modifiers as part of the entity in order to ensure a coherent part-whole pair. To this end, we asked annotators to tag not only the head, but also any surrounding text which was essential to the meaning of the head.

Unfortunately, we found that this distinction was very difficult to make in practice. The annotators struggled with many ambiguous instances, and we struggled to formulate a coherent and consistent standard for when to mark what. While the toy examples presented in our guideline were clear, the actual corpus had a very wide variety of potentially important modifiers. If we were to capture them, we would likely be trading away a significant amount of inter-annotator agreement.

Instead, we decided to always annotate only the head of the noun phrase. While some detail is lost by this change, we felt it was more important to be able to give the annotators and clear and concise standard to follow. Further, it may be possible to recover many of these lost modifiers in postprocessing if necessary.

7 IAA

Each document in our corpus (with a few exceptions) was annotated twice. We used these pairs of duplicate documents to calculate Cohen's Kappa as a measure of inter-annotator agreement on our extents and attributes. Cohen's Kappa is calculated using two terms - $Pr(a)$, which is the observed rate of agreement, and $Pr(e)$, which is the rate of agreement expected by chance. The calculation of $Pr(e)$ is somewhat complicated, as a truly accurate measurement would require knowing the exact probability of a particular annotator marking each word, which is of course not consistent across words (all annotators are more likely to mark nouns as entities than they are prepositions, for example). However, since this sort of information is obviously unavailable, we made the simplifying assumption that each word in a document was in fact equally probable to be marked. That probability was taken to be the total number of words in the document divided by the number of words actually tagged by a given annotator.

We also allowed for somewhat generous matching when deciding if two annotators had marked the same extent. While the most rigorous method would be to count a match only if the two extents had exactly the same start and end offsets, this would exclude many cases in which annotators accidentally included a space or punctuation in their extents. Instead, we count any pair of overlapping extents to be a match.

Across our documents, we saw a wide range of kappa scores for both extents and attributes. The scores for ENTITY agreement are shown below:

File	ENTITY Kappa
khoisan languages	-0.08
tynwald	-0.07
statistics	-0.02
prudential center	-0.01
mathematics	0.00
computer science	0.00
ark of the covenant	0.07
paint	0.11
plot	0.13
parliamentary system	0.16
spotify	0.17
republics of russia	0.18
madden nfl 08	0.22
republic	0.25
tarbell building	0.30
english	0.32
single market	0.36
rock music	0.38
bojjannakonda	0.38
chordate	0.38
gravel road	0.39
computing	0.41
algorithm	0.41
ore	0.43
chemical element	0.44
insect	0.49
palaeeudyptes	0.49
physics	0.51
telephone	0.53
sea snail	0.53
computer network	0.59
city	0.59
atomic number	0.73
yacht	0.79
motorcycle speedway	0.79

Although there is some disagreement about how to interpret Cohen's Kappa, Landis & Koch(9) suggested that a score < 0 indicates no agreement, $0 - .2$ indicates slight agreement, $.2 - .4$ fair agreement, $.4 - .6$ moderate agreement, $.6 - .8$ substantial agreement and $.8 - 1$ nearly perfect agreement. The table is divided into sections to indicate these ranges.

As can be seen, the majority of our documents fell into the fair to moderate range, with many documents clustered around .4. We also had a few documents with strong agreement, and a significant minority with poor agreement.

We see that, with some exceptions, the most successful files are those dealing with physical items, often with mechanical or anatomical parts. Files like yacht, sea snail and insect had many very clear meronymic relationships dealing with this sort of physical objects. On the other hand, files dealing with more abstract concepts, such as language, government and general fields of study saw much poorer agreement.

We suspect that this is because the physical part-whole relationship is the most familiar, where as abstract relationships are more distant conceptually. As such, annotators were able to use their existing skill at identifying the physical relationships to do so more consistently. The poorer results on abstract concepts suggests a need for improvement of our guideline with respect to this sort of example.

Additionally, there can be a tendency when annotating to look for more examples to tag when a file seems sparse. The natural assumption when given a file would be that there are some good examples of the target phenomenon in that file. If none or few are found, the annotator may expand their definition to fit ambiguous cases that otherwise would have been rejected. Because documents about abstract concepts often have many fewer meronyms than those about physical objects, this tendency may have impacted results for abstract topics more severely.

Another factor which may have hurt agreement on some files was our exclusion of some types of meronyms. For example, many sources consider classifications to be meronyms (such as "Wolves belong to the genus *canis*"). We, however, exclude this relationship entirely. This means that if annotators find an ambiguous case between this type of relationship and a member-group relationship, they may simply decline to mark it entirely. The same is true for other types of meronyms which we exclude, such as geographic meronyms (which can be ambiguous when dealing with a noun that is both a location and a political entity), and activity-based meronyms (which can be ambiguous in fields of study, which can be viewed both as an activity and a body of knowledge).

It may be worth reconsidering the decision to exclude these other types of meronyms, because including them would relegate much of the ambiguity to the type attribute instead of the extent itself. This would give us a clearer picture of the top-level phenomenon of meronyms as a whole, and make it easier to resolve disagreements over type.

The SIGNAL tag showed similar patterns of agreement to the ENTITY tag:

File	SIGNAL Kappa
statistics	-0.01
spotify	-0.01
prudential center	0.00
khoisan languages	0.00
mathematics	0.00
computer science	0.00
ark of the covenant	0.08
parliamentary system	0.11
paint	0.13
chordate	0.18
plot	0.18
rock music	0.22
insect	0.23
republic	0.23
algorithm	0.25
tarbell building	0.25
english	0.28
republics of russia	0.33
madden nfl 08	0.39
physics	0.39
palaeeudyptes	0.40
city	0.44
bojjannakonda	0.45
telephone	0.45
computing	0.45
ore	0.49
gravel road	0.50
tynwald	0.50
chemical element	0.55
single market	0.59
computer network	0.62
motorcycle speedway	0.66
atomic number	0.66
sea snail	0.69
yacht	0.74

It should not be surprising that the agreement for SIGNAL closely follows the agreement for ENTITY, because a SIGNAL word can only be present when a meronym is identified. Agreement is generally slightly lower, because agreement on a SIGNAL word requires agreement on both the underlying relationship and which word signals it.

We also calculated agreement for the attributes of the ENTITY tag (type and reference count). When we include cases on which anntotators disagree about the underlying extents, the results relatively closely follow the scores for ENTITY. To get a clearer picture of the agreement on these attributes alone, we considered only the subset on which annotators agreed on the underlying extent. Obviously, this requires dropping some of our tagged ENTITIES, and so may not be completely representative of the data set, but we feel it gives a better understanding of the performance on the attributes alone. Note that a file is excluded entirely if no ENTITY tags were agreed upon.

File	TYPE Kappa
tynwald	-1.17
insect	-1.17
republic	-1.14
republics of russia	-0.33
computing	-0.18
spotify	-0.13
algorithm	-0.03
telephone	0.03
chemical element	0.03
single market	0.19
computer network	0.2
yacht	0.36
rock music	0.71
ark of the covenant	1
plot	1
sea snail	1
gravel road	1
motorcycle speedway	1
palaeeudyptes	1
atomic number	1
chordate	1
english	1
physics	1
city	1
ore	1
parliamentary system	1
paint	1
bojjannakonda	1
madden nfl 08	1
tarbell building	1

Many files had perfect agreement on whether an entity was physical or abstract, but a surprising number had very poor agreement. There are a few contributing factors to this. One is the danger of default values for attributes and another is our lax treatment of these attributes in our guidelines.

When designing our DTD, we noted that most entities in our corpus were physical, and so made that the default value of the type attribute. However, this can lead annotators to forget to change the default when dealing with the rarer abstract case, especially in a file with many physical entities and few abstract ones. This is exacerbated by the fact that identifying entity characteristics is not the main focus of the task, and so is more easily overlooked.

Many files had perfect agreement on whether an entity was physical or abstract, but a surprising number had very poor agreement. There are a few contributing factors to this. One is the danger of default values for attributes and another is our lax treatment of these attributes in our guidelines.

When designing our DTD, we noted that most entities in our corpus were physical, and so made that the default value of the type attribute. However, this can lead annotators to forget to change the default when dealing with the rarer abstract case, especially in a file with many physical entities and few abstract ones. This is exacerbated by the fact that identifying entity characteristics is not the main focus of the task, and so is more easily overlooked. One option to alleviate this problem would be to integrate preprocessing to determine likely values for the attribute, and alert the annotator if a mismatch was found between these values and their annotation. However, this would require significant modifications to the annotation environment.

Because this attribute was not part of the phenomenon which we are trying to identify, our explanation of how to properly annotate it was somewhat brief in our guidelines. As can be seen by our Kappa scores, this turned out to be a task with many non-trivial cases, and we would have done well to provide a more thorough explanation of it to our annotators. We underestimated its complexity, and assumed it would be quite easy to annotate.

In fact, there are many cases where a noun can have both physical and abstract interpretations. For example, in our document about Khoisan languages, we see Africa used to refer to both the geographic location (a physical entity) and the cultural amalgamation of its people (an abstract entity). This sort of double meaning contributed to many of the poor kappa scores.

Our scores for the reference count attribute saw very similar kappa scores, and we suspect the same causes. Our guideline again provided only a brief explanation, but we later encountered significantly more complex examples in our corpus, especially when dealing with the mass noun and zero count values.

8 Machine Learning

8.1 Problem Formulation

We formulate part-whole relation extraction as a simple classification problem. The first step is to detect all entities within in a document that could participate in a part-whole relationship. In the corpus, this is done by simply using the annotated entities while disregarding other information, while for unlabed corpora, this would be done by noun or noun-phrase chunking. After all of the entities are found, each pair is considered to potentially participate in a part-whole relationship. Since most part-whole relationsips occur within a single sentence or adjacent sentences, a sentence distance could (and should) be imposed, but we did not do that for our results. Then, for each pair of entities in the document, we create a 'sample' for the machine learnign that consists of features for the individual entities and the combination of entities and the class label is a binary "True" or "False" corresponding to the presence or absence of a part-whole relationship respectively.

8.2 Features

From the beginning, we wanted to choose features that were, like prior works, based on syntax and patterns. All of our features were derived using the NLTK (Natural Language Tool Kit) package for Python and Java MaltParser program. The NLTK provides a wide range of natural language processing functions, some of which are described here and the MaltParser is a library for parsing sentences into their dependency tree structures. Our first feature was part of speech, which we derived using NLTK, after also using NLTK to split the raw Wikipedia articles into sentences, then tokens. Part of speech is an important feature because, in general, all entities are nouns, but even amongst nouns, there are few forms of nouns and the specific subtypes give subtle clues to the existence or lack-there-of of part-whole relations. The second feature is the entity's dependency role. This feature and all other dependency tree related features were extracted by running and parsing the output of the MaltParser from within Python. Intuitively, the dependency role is important because it describes the role played by the entity in some relationship that the sentence describes. The next features are dependency tree path distance between entities and the dependency tree path distances to the root of the dependency tree. These are useful because they describe how "close" their dependency relationships are and how close each entitie's role is to the main or overall information the sentence conveys. We also included slight variations of the above numerical features, particularly the floored log (base 2) since the distribution of tree distances in our corpus tends to follow a logarithmic distribution. This also helps counteract our small corpus size.

8.3 Results

On top of the tools used to perform feature extraction, we used Weka 3.6.9 to perform the actual classifications. Weka is an open-source collection of many popular classifiers that allows for a great deal of classifier-specific options and automation. One weakness of Weka is it's inability to handle large data sets such as ours. Though we had few part-whole relationships in our corpus, totaling at 525, a very large number of negative samples were created after pairing every entity in each document with every other entity in that document. The total number of samples was over 55,000. Because of the limitations of Weka we ran a few of the faster, well-known, and popular classifiers and summarize the results below showing various statistics (true positive rate, false positive rate, precision, recall, and f1-measure) and confusion matricies. As with prior work, our main focus was to get the true positive rate for the True class (indicating a part-whole relationship) as large as possible. This true positive rate corresponds to the ratio of samples that are correctly recognized as part-whole's among all of the part-whole samples. We found that Naive Bayesian classifiers outperform the other classifiers used by a surprising 40% with respect to the true positive rate of True samples.

8.3.1 Naive Bayes (Updatable)

TP	FP	Prec.	Recall	F1	Class
0.969	0.391	0.996	0.969	0.982	False
0.609	0.031	0.156	0.609	0.249	True
0.966	0.387	0.988	0.966	0.975	

False	True	← (classified as)
52936	1698	False
202	315	True

8.3.2 Bayes Network

TP	FP	Prec.	Recall	F1	Class
0.969	0.383	0.996	0.969	0.982	False
0.617	0.031	0.159	0.617	0.252	True
0.966	0.38	0.988	0.966	0.976	

False	True	← (classified as)
52943	1691	False
198	319	True

8.3.3 Voted Perceptron

TP	FP	Prec.	Recall	F1	Class
1	0.992	0.991	1	0.995	False
0.008	0	0.308	0.008	0.015	True
0.991	0.983	0.984	0.991	0.986	

False	True	← (classified as)
54625	9	False
513	4	True

8.3.4 Random Forest

TP	FP	Prec.	Recall	F1	Class
0.998	0.793	0.993	0.998	0.995	False
0.207	0.002	0.476	0.207	0.288	True
0.99	0.786	0.988	0.99	0.989	

False	True	← (classified as)
54516	118	False
410	107	True

8.4 Future Work

For machine learning, the core of future work is to add more features, especially features based on syntax trees, and to balance the number of positive and negative samples.

Some key features we'd like to add are the number of occurrences of signal words identified in the annotated data near or between two entities, the syntax tree path distance, the dependency head of each entity, and a boolean for whether the entity heads match. The signal word count should help the classifier because these are words that have been manually identified as indicators of meronymic relationships. Syntax tree path distance, and other syntactic features, can contribute to machine learning because the relationships are expressed syntactically and syntax tree distance is generally correlated with expressed relationships in a sentence. Finally, the dependency head of the entities directly indicate *a relationship* that the sentence *explicitly* states the entities participate it and what type of relationship that is. If this is a part-whole relationship, then we could rely on that feature alone, and if the two heads match, that means that sentence explicitly states that they participate in the same relationship.

To address balancing the corpus, there are two general approaches; removing negative samples and 'duplicating' positive samples (in the training data). Removing negative samples is straightforward and can be done by simply removing a random subset of the negative samples or ruling out 'obvious' negative samples such as entities that have a large sentence distance. In general, part-whole relationships that are directly expressed (like the ones we are interested in for machine learning) occur within a sentence or two adjacent sentences. Duplicating samples is similarly straightforward, though we came up with a method we call "fuzzing" that may be more beneficial than duplicating the samples exactly. Fuzzing consists of taking numerical attributes, such as tree distances, and adding a small amount, such as -1 or 1. Since numerical features such as tree distance should have some tolerance, we believe this strategy will allow a small corpus like ours to cover a larger feature space and give more positive samples, thereby improving classification, while not harming classification since we expect that the unannotated data the classifier would be used on would contain small variations that do not indicate a change in the presence or absence of part-whole relationships.

9 Future Work

9.1 Anaphora

A potential route for expanding the specification is the handling of anaphora. Because we tag the entity which is most directly syntactically involved in the meronymic relationship, we often tag anaphora instead of the original referent object. In the case of lengthy descriptions of an object's constituent parts, this can often mean that the original referent is appears several sentences prior.

On the one hand, knowing that "it" is part of "that" is much less helpful than having the referents resolved, so we would like to have a way of indicating what those referents are. If a corpus has many instances of anaphora, even an otherwise strong machine learning algorithm that does not account for resolving them will produce many useless results like the example above.

However, the alternative of simply tagging the original referent as the entity creates a much more difficult learning environment, because those referents may be far removed from the syntactic clues to the existence of the meronym. Without some understanding of anaphora, a machine learning algorithm would be able to make little use of an example where one entity is separated by a full paragraph from the other.

Unfortunately, anaphora resolution is a non-trivial problem, so introducing it alongside meronymy detection would greatly complicate the task. We would, in effect, be asking the computer to learn two complex phenomena at once, as well as their interaction. So, while a successful system which handled anaphora would almost surely produce much better results than an otherwise successful system which did not, the cost of that extension would be quite significant. As a result, we opted to keep the task focused on meronyms alone.

9.2 Additional types

Another avenue for expansion, as mentioned earlier, would be to handle other types of meronyms which we originally excluded for simplicity. Some of these would be easily incorporated into the existing frame work, such as geographic meronyms and classification meronyms, which have roughly the same structure and components as our current types.

Other types, especially activity-based meronyms, would require changes or additions to our specification. Whereas the existing types use nouns as entities, these meronyms often have verbs as their constituent parts. This means the addition of an entirely new part of speech to our specification, along with new attributes for these verbal entities.

As we expand the types of meronyms covered, the machine learning component may become more difficult, as each type will have its own common structure and signals, which may differ significantly from the others. This might muddle the feature set and make learning less successful.

9.3 Robust Signals

Finally, we might consider a more robust handling of signals. Presently, we focus on words as the only form of signal. However, sometimes it is only a part of a word that signals for a meronymic relationship, as in the case of possessives. While it is possible to simply tag the entire possessive word as a signal, it would be more accurate to view the fact of the possessive form as the signal, rather than the word which is in that form. This could be accomplished by adding attribute(s) to the SIGNAL tag to allow the annotator to specify what about the word makes it a signal.

Similarly, we sometimes see adjectives that indicate a meronymic relationship, especially for the substance type (for example, "a wooden bat" or "a golden ring"). While these cases clearly indicate a substance relationship, they don't fit well into our current specification and guideline, which want entities to be nouns rather than adjectives. A change to handle adjectives would also incorporate the above concept of word form as a signal, because the fact of the adjective form is serving as the signal, but the underlying root is serving as the entity.

A third unconventional form of signal is word compounding, where neither word is itself a signal, but rather their placement next to each other. For example, in the compound “tuna salad”, we can tag tuna and salad as entities, but lack a way to indicate that their relationship is signaled by compounding. This type of signal may be especially difficult to capture using extent-based annotation, because no particular string of text is acting as the signal. Handling it effectively may require a change to the annotation paradigm which can sensibly handle signals as structures instead of just as extents.

10 Conclusion

Overall, our guideline and specification saw moderate success in enabling our annotators to accurately capture our target phenomenon. While we were able to acheive strong agreement for certain domains (physical objects in traditional part-whole relationships), our guideline fell short for more complicated domains, especially those involving abstract objects. However, we feel that our areas of success indicate the potential of significant improvement in those areas where we struggled. With revisions to the guideline, and potential revisions to the annotation procedure, we believe strong agreement on all domains would be possible.

Our machine learning efforts, though still preliminary, show promising results. We demonstrate an improvement over baseline random chance, and our annotation enables a richer feature set than simple bag of words and n-grams. Further work on enriching the feature set, corpus and annotation would be expected to provide additional gains in classification accuracy.

The task of identifying and classifying meronymic relationships presents is non-trivial, and presents many challenges for both annotation and learning. While some of these challenges remain to be overcome, our current work and results form a foundation for future work, as well as valuable lessons-learned for building a more effective annotation.

References

[1] Alessandro Artale, Enrico Franconi, Nicola Guarino *Open problems with part-whole relation.* Applied Ontology 0 (2007)

[2] Ulrike Sattler *A concept language for an engineering application with part-whole relations.* Proceedings of the International Workshop on Description Logics, pages 119-123

[3] Uta E. Priss, *The formalization of WordNet by Methods of Relational Concept Analysis.* 1996.

[4] Cruse, D. A. *Lexical Semantics* Cambridge, New York.

[5] Iris, Madelyn; Litowitz, Bonnie; Evens, Martha *Problems of Part-Whole Relations.* In: Evens, Martha W. (ed.). Relational Models of the Lexicon. Cambridge University Press.

[6] Morton E. Winston, Roger Chaffin, Douglas Herrmann, *A Taxonomy of Part Whole Relation.* Cognitive Science, Volume 11, Issue 4, pages 417–444, October 1987

[7] C. Maria Keet, Alessandro Artale *Representing and Reasoning over a Taxonomy of Part-Whole Relations.* Applied Ontology 0 (2007)

[8] Roxana Girju, Dan Moldovan, Adriana Badulescu *Automatic Discovery of Part–Whole Relations.* Journal Computational Linguistics archive, Volume 32 Issue 1, March 2006, Pages 83-135

[9] Landis, J.R.; & Koch, G.G. *The measurement of observer agreement for categorical data.* Biometrics 33 (1), 1977, Pages 159–174.

[10] Robert, Angus. *Learning Meronyms from Biomedical Text.* In Proceedings of the ACL Student Research Workshop (June 2005), pp. 49-54