A Method for Learning Part-Whole Relations

Willem Robert van Hage[1,2], Hap Kolb[1], and Guus Schreiber[2]

[1] TNO Science & Industry Delft, wrvhage@few.vu.nl, hap.kolb@tno.nl
[2] Vrije Universiteit Amsterdam, schreiber@cs.vu.nl

Abstract. Part-whole relations are important in many domains, but typically receive less attention than subsumption relation. In this paper we describe a method for finding part-whole relations. The method consists of two steps: (i) finding phrase patterns for both explicit and implicit part-whole relations, and (ii) applying these patterns to find part-whole relation instances. We show results of applying this method to a domain of finding sources of carcinogens.

1 Introduction

A plethora of existing vocabularies, terminologies and thesauri provide key knowledge needed to make the Semantic Web work. However, in using these sources witinin one context, a process of alignment is needed. This has already been identified as a central problem in semantic-web research. Most aligment approaches focus on finding equivalence and or subclass relations between concepts in diffeent sources. The objective of this paper is to identifying alignment relations of the part-whole type. Part-whole relations play a key role in many application domains. For example, part-whole is a central structuring principle in artefact design (ships, cars), in chemistry (structure of a substance) and medicine (anatomy). The nature of part-whole has been studied in the area of formal ontology (*e.g.*, [1]). Traditionally, part-whole receives much less attention than the subclass/subsumption relation.

The main objective of this paper is to develop a method for learning part-whole relations from existing vocabularies and text sources. Our sample domain is concerned with food ingredients. We discuss a method to learn part-whole relations by first learning phrase patterns that connect parts to wholes from a training set of known part-whole pairs using a search engine, and then applying the patterns to find new part-whole relations, again using a search engine. We apply this method in a use case of assisting safety and health researchers in finding sources of carcinogenic substances using Google. We evaluate the performance of the pattern-learning and the relation-learning steps, with special attention to the performance of patterns that implicitly mention part-whole relations. Furthermore we perform an end-to-end task evaluation to establish whether our method accomplishes the task.

In Sec. 2 we describe the use case on which we evaluate end-to-end performance and pose performance criteria. In Sec. 3 we discuss the experimental set-up we use to learn part-whole relations. In Secs. 4 and 5 we describe the learning and application of patterns to find part-whole relations and evaluate the performance of the patterns in terms of Precision. In Sec. 6 we evaluate Recall on four sample carcinogens. Sec. 7 discusses related work. We conclude with a discussion of the results and open research questions in Sec. 8.

2 Use Case

An important application area of part-whole learning is health and safety research. Experts in this field are faced with hard information retrieval tasks on a regular bases. News of a benzene spill in a river, for example, will trigger questions like "Is the general public's health in danger?", "Are there any foodstuffs we should avoid?", and "Are there any occupational risks, fishermen perhaps?". The first task the health and safety researchers are faced with is to find out via which pathways the substance in question can reach humans. Only then can they investigate if any of these pathways apply to the current situation. A sizable part of this problem can be reduced to finding all part-whole relations between the substance and initially unknown wholes in scientific literature and reports from authorities in the field such as the United States Food and Drugs Administration[3] (FDA) and Environmental Protection Agency[4] (EPA), and the World Health Organization[5] (WHO).

The wholes should be possible routes through which humans can be exposed to the substance. For example, tap water, exhaust fumes, or fish. We will not go into detail discussing the roles these concepts play that leads to the actual exposure. For example, when humans are exposed to benzene in fish by eating the fish, fish assumes the role of food. Relevant part-whole relations can be of any of the types described by Winston, Chaffin, and Herrmann [12].

component/integral object "Residents might have been exposed to *benzene* in their *drinking water*."
member/collection "*Benzene* belongs in the group of *BTX-aromatics*."
portion/mass "*3 tons* of the *benzene emissions* can be attributed to the dehydrator."
stuff/object "*Aftershave* used to contain *benzene*."
feature/activity "*Benzene* is used in the *dehydration process*." The part in this case is not benzene itself, but the application of benzene, which is abstracted over with the word "used".
place/area "*Benzene* was found in the *river*." The part in this case is the location where the benzene was found, which is left anonymous.

The automation of the knowledge discovery task described above is a success if and only if the following criteria are met:

1. The key concepts of each important pathway through with a carcinogen can reach humans should be found. (*i.e.,* Recall should be very high.)
2. The researchers should not be distracted by too many red herrings. (*i.e.,* Precision should be sufficient.)

Precision can be evaluated in a straightforward manner by counting how many of the returned part-whole relations are valid. The evaluation of Recall however poses a greater problem. We are attempting to learn unknown facts. How can one measure which percentage of the unknown facts has been learnt when the facts are unknown? For this

[3] `http://www.fda.gov`
[4] `http://www.epa.gov`
[5] `http://www.who.int`

use case we will solve this problem by looking at exposure crises for four substances (acrylamide, asbestos, benzene, and dioxins) that have been documented in the past. We know now which pathways led to the exposure in the past. This means we can construct sets of pathways we should have known at the time of these crises and use these sets to evaluate Recall.

3 Experimental Set-up

In this paper we will use two-step method to learn part-whole relations. First we learn lexical patterns from known part-whole pairs, using search engine queries. Then we apply these patterns to a set of parts to find wholes that are related to these parts, also using search engine queries. To constrain the size of the search space we will constrain both the set of parts and the set of wholes to controlled vocabularies. In more detail, the method works as follows:

1. **Learning part-whole patterns.**
   (a) Construct a search query for each part-whole pair in a training set.
   (b) Collect phrases from the search results that contain the part-whole pair.
   (c) Abstract over the parts and wholes in the phrases to get patterns.
   (d) Sort the patterns by frequency of occurrence. Discard the bottom of the list.
2. **Learning wholes by applying the patterns.**
   (a) Fill in each pattern with all parts from a set of part instances, while keeping the wholes free.
   (b) Construct search queries for each filled in pattern.
   (c) Collect phrases from the search result that contain the filled in pattern.
   (d) Extract the part-whole pairs from the phrases.
   (e) Constrain the pairs to those with wholes from a controlled vocabulary.
   (f) Sort the pairs by frequency of occurrence. Discard the bottom of the list.

In the following two sections we will describe the details of the data sets we used and we will motivate the decisions we made.

4 Learning Part-Whole Patterns

In this section we will describe the details of step 1 in our part-whole learning method, described in the previous section. We will describe the training set we used and the details of the application of step 1 on this training set, and analyze the resulting patterns.

Our training set consists of 503 part-whole pairs, derived from a list of various kinds of food additives and food product types they can occur in created by the International Food Information Council[6] (IFIC) and the FDA.[7] The list contains 58 additives (parts) and 113 food products (wholes), grouped together in 18 classes of additives such as sweeteners and preservatives. An example is shown in Fig. 1. It is not specified which additives occur in which food products. To discover this, we took the cartesian product of the additives and the food products and filtered out the pairs that

[6] `http://www.ific.org`
[7] `http://www.cfsan.fda.gov/~dms/foodic.html`

yielded no hits on Google[8] when put together in a wildcard query. For example, the pair ⟨table-top sugar, aspartame⟩ is filtered out, because the query `"table-top sugar * aspartame" or "aspartame * table-top sugar"` yields no hits.

Type	Sweeteners
What They Do	Add sweetness with or without the extra calories.
Examples of Uses	Beverages, baked goods, confections, table-top sugar, substitutes, many processed foods.
Product Label Names	Sucrose (sugar), glucose, fructose, sorbitol, mannitol, corn syrup, high fructose corn syrup, saccharin, aspartame, sucralose, acesulfame potassium (acesulfame-K), neotame

Fig. 1. An excerpt from the IFIC and FDA list of food additives.

For all 503 part-whole pairs that did yield results we collected the first 1000 snippets (or as many snippets as were available). We attempted to part-of-speech tag these snippets. This did not produce good results, because nearly all snippets were incomplete sentences and many were lists of substances. For example, "... Water)*, Xanthan Gum, Brassica Campestris (Rapeseed), Essential Oils [+/- CI 77491,CI ...". None of the part-of-speech taggers we tried were able to deal with this. Therefore we used the untagged snippets and looked up all consistent phrases that connected the part and whole from the query. In these phrases we substituted all parts and wholes by the variables "part and whole". This yielded 4502 unique patterns, which we sorted by frequency of occurrence. The frequencies of the patterns are shown in Fig. 2.

Due to the fact that there were many lists of substances in our data there were also many patterns that did not describe a part-whole relation, but that were merely part of a list of substances containing the part and the whole. These patterns can be easily recognized, because they contain names of substances. For example, for the pair ⟨cheese, enzymes⟩ the following snippet was returned: "*cheese* (pasteurized milk, cheese cultures, salt, *enzymes*)". An example of a good snippet is: "All *cheese* contains *enzymes*.". To exclude lists we removed all patterns that contain, apart from the part and whole, labels of concepts in agricultural thesauri. The thesauri we used are the NAL Agricultural Thesaurus[9] and the AGROVOC Thesaurus[10]. (We used the SKOS[11] version of these thesauri.) This filtered out 1491 patterns, of which only 12 were correct part-whole patterns. Fig. 2 shows a Precision graph of the list of patterns before and after the filtering step.

To restrict the number of Google queries needed to find wholes for parts we decided not to use all of the remaining 3011 patterns, but to select the most productive patterns. We analyzed the 300 patterns that produce the most results. For each pattern we looked at the snippets it returned. If the majority of the occurrences of the pattern described a

[8] http://www.google.com
[9] http://agclass.nal.usda.gov/agt
[10] http://www.fao.org/agrovoc
[11] http://www.w3.org/2004/02/skos

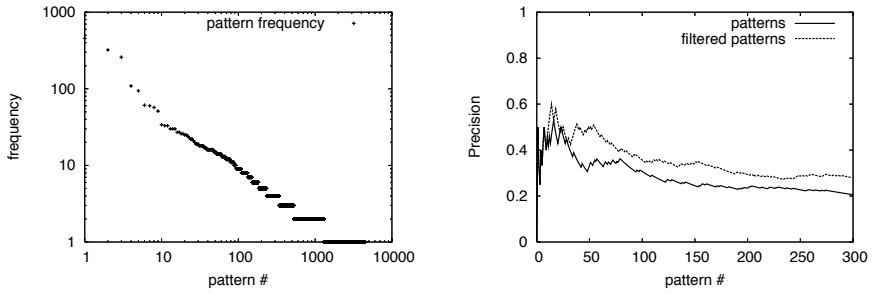

Fig. 2. (left) Frequency distribution in the training set of the learnt patterns. Referred to as T in Table 3. (right) Precision@n (*i.e.*, # correct part of patterns in the top-n / n) graph over the top-300 most frequent patterns, before and after filtering out patterns that contain labels of AGROVOC or NALT concepts.

proper part-whole relation (*i.e.*, Precision $\geq$ .5) we classified the pattern as part-whole. Otherwise we classified it as not part-whole.

We distinguished the following groups of patterns, based on the most common types of errors that led to the classification of the pattern as not part-whole. A pattern can yield more than one type of false relations, but the classification is based on the most common of the error types.

too specific Too training-set specific to be useful. Either the pattern contains adjectives or it yields no hits due to over-training.

too generic The pattern matches part-whole relations, but also too many non-part-whole relations to be useful. For example, the pattern "whole part", as in "barn door", can match any type of collocation.

is a The pattern primarily matches hyponyms. The language used to describe member/collection relations is also used for hyponyms.

conjunction/disjunction The pattern primarily matches conjunctions / disjunctions.

related The pattern connects terms that are related, but not part-whole related.

wrong Not a proper pattern for any other reason. Most of the errors in the wrong category can be attributed to the lack of sophisticated linguistic analysis of the phrases.

Table 2 shows the build-up of the different error types.

We corrected 6 patterns that were classified as not part-whole, and added them to the part-whole patterns. These patterns are not counted in Table 2. They are listed in Table 1. Notice that in the English grammar, hyphenation turns a part-whole relation into its inverse. For example, "sugar-containing cake" and "cake containing sugar".

While analyzing the correct part-whole patterns we noticed that the phrases that deal with part-whole relations do not always explicitly state that relation. Often, the part-whole relation has to be inferred from the description of a process that led to the inclusion of the part in the whole or the extraction of the part from the whole. For example, from the sentence "I add *honey* to my *tea*." we can infer that honey is part of

"part to whole"	→ "add part to whole", "added part to whole"
"part to the whole"	→ "add part to the whole", "added part to the whole"
"part gives the whole"	→ "part gives the whole its"
"part containing whole"	→ "part-containing whole"
"part reduced whole"	→ "part-reduced whole"
"part increased whole"	→ "part-increased whole"

Table 1. Manually corrected patterns.

the tea, even though the sentence only mentions the process of adding it. In addition to explicit descriptions of part-whole relations we distinguish two types of phrases that mention part-whole relations implicitly.

part of The phrase explicitly describes a part-whole relation. For example, "There's alcohol in beer.".

source of The phrase implicitly describes a part-whole relation by describing the action of acquiring the part from the whole. For example, "Go get some *water* from the *well*.".

made with The phrase implicitly describes a part-whole relation by describing a (construction) process that leads to a part-whole relation. For example, "I add *honey* to my *tea*".

Table 2 shows that together, the implicit patterns account for a third of the total number of part-whole pairs.

When applying patterns to learn part-whole relations it is useful to make this distinction into three types, because it turns out that these three types have rather different Precision and Recall properties, listed in Table 3. The patterns in the part of class yield the most results with high Precision. The patterns in the made with class also yield many results, but—somewhat surprisingly—with much lower Precision, while the patterns in the source of class yield few results, but with high Precision.

The 91 patterns we used for the discovery for wholes are the 83 classified as part-whole in Table 2 and the 8 listed in Table 1 on the right side. They are listed in Table 6.

5 Finding Wholes

In this section we will describe the details of step 2 in our part-whole learning method, described in the previous section. We will describe the sets of part and whole instances we used, and analyze the resulting part-whole relations.

In the use case we focus on finding wholes that contain a specific substance. Initially, any concept name is a valid candidate for a whole. We tackle this problem by first reducing the set of valid wholes to those that occur in a phrase that matches one of the patterns learnt in step 1 of our method. This corresponds to step 2c and 2d of our method. Then we prune this set of potential wholes using two large, agricultural, and environmental thesauri that are geared to indexing documents relevant to our use

pattern class	example pattern	# patterns in class
part-whole		83
part of	whole containing part	40
made with	part added to whole	36
source of	part found in whole	7
not part-whole		217
wrong	part these whole, part organic whole	186
too specific	part in commercial whole	10
too generic	part of whole	7
is a	whole such as part	5
related	part as well as whole	4
conjunction	part and whole, whole and part	3
disjunction	part or whole, whole or part	2

Table 2. Analysis of the top-300 most frequently occurring patterns.

case. We remove all wholes that do not match a concept label in either thesaurus. This corresponds to step 2e of our method. The former reduction step asserts that there is a part-whole relation. The latter that the whole is on topic.

We select the possible part instances from a list of carcinogens provided by the International Agency for Research on Cancer[12] (IARC). In the IARC Monographs on the Evaluation of Carcinogenic Risks to Humans[13] carcinogenic agents, mixtures and exposures are classified into four groups: positively carcinogenic to humans, probably or possibly carcinogenic to humans, not classifiable as carcinogenic to humans, and probably not carcinogenic to humans. We took the agents and mixtures from the group of positively carcinogenic factors. We interpreted each line in the list as a description of a concept. We removed the references and expanded the conjunctions, interpreting each conjunct as a label of the concept. i.e., For example, we transform the list entry "Arsenic [7440-38-2] and arsenic compounds (Vol. 23, Suppl. 7;1987)" into a concept arsenic with the labels "Arsenic" and "arsenic compounds". The resulting list contains 73 concepts, with 109 labels in total. We applied the 91 patterns that resulted from the process described Sec. 4 on these 109 labels to discover wholes. We allow for words—generally articles and adjectives—to appear in between the whole and the rest of the pattern. For example, the pattern "part in whole" can be interpreted as "part in $*$ whole", and hence will match "part in deep-sea whole" and "part in the whole". This also means there can be overlap between the sets of part-whole pairs retrieved by patterns. From the resulting filled-in patterns we extracted the wholes. We filtered out all wholes from this list that do not appear in the UN FAO AGROVOC Thesaurus and the USDA NAL Agricultural Thesaurus. When put together, these thesauri contain 69,746 concepts with 87,357 labels in total. Thus limiting the set of discoverable wholes to 69,746 concepts. For each remaining whole in the list we construct a part-whole relation.

An assessment of the part-whole results is shown in Table 6. We approximated Precision for the 91 patterns we used to find wholes based on a random sample of

[12] http://www.iarc.fr

[13] http://monographs.iarc.fr/ENG/Classification

25 discovered pairs. The results are shown under "Precision". The number of hits per pattern are listed under *D*. This number includes duplicate phrases and multiple phrases describing the same part-whole pair. Table 4 in Sec. 6 shows how many unique wholes are found for four example parts.

pattern class	# patterns in class	*T*	*D*	avg. Precision
part of	40	744	84852	.81
made with	36	525	33408	.69
source of	7	111	8497	.83

Table 3. Average pattern performance per pattern class. *T* is the number of times patterns in the class occur in the training set. *D* is the number of discovered part-whole phrases.

6 Analysis

In Sec. 2 we stated two criteria that have to be met for the application of our part-whole learning method to be a success. Precision has to be sufficient, and Recall has to be very high. In Secs. 4 and 5 we analyzed the results in terms of frequency and Precision. We achieved an average Precision of .74. In this section we will assess Recall.

Since even the knowledge of experts of whether or not a substance is contained in some whole is far from complete we can not create a complete gold standard to measure Recall. It is simply infeasible. We can, however, approximate Recall by computing it on samples.

We set up four test cases centered towards discovering possible causes of exposure to a specific carcinogenic agent. The agents we chose are acrylamide, asbestos, benzene, and dioxins. These substances have all caused health safety crises in the past and possible exposure to them has been extensively documented. For each case we decided on 15 important concepts that contain the carcinogen and define a possible exposure route. For example, you can be exposed to acrylamide by eating fried food such as french fries, because acrylamide can be formed in the frying process. The selection of the wholes was based on reports from the United States Environmental Protection Agency (EPA) and the Netherlands Organization for Applied Scientific Research (TNO) Quality of Life. The cases were set up without knowledge of the data set and the learning system, to minimize the hindsight bias, but with knowledge of the concepts in the AGROVOC and NALT thesauri. The sets of wholes are shown in Table 5, along with the rank at which the whole occurs in the list of discovered wholes. Recall and the total number of discovered wholes are shown in Table 4.

For all of the cases we found a large majority of the important concepts. For half of the missed concepts we found concepts that are very closely related. For example, we did not find the concept "cement pipes", but we did find "cement" and "pipes", and we did not find "air", but we did find "air pollution" and "atmosphere".

The data sets and the results can be found at the following web location: `http://www.few.vu.nl/~wrvhage/carcinogens`.

concept (part)	# of wholes found	Recall
acrylamide	350	13/15 (.86)
asbestos	402	11/15 (.73)
benzene	479	13/15 (.86)
dioxins	439	12/15 (.80)

Table 4. Recall on four sample substances.

7 Related Work

The method of automatic learning of relations by first learning patterns and then applying these patterns on a large corpus is widely used. An example in the domain of business mergers and production is described in the 1999 article by Finkelstein-Landau and Morin [5]. Their work on extracting companies-product relations touches lightly upon the subject of this paper. Another example of pattern-based relation learning on the web is the KnowItAll system of Etzioni et al. [4]. The learning of part-whole relations however is quite rare. Two examples, are the work of Berland and Charniak in 1999 [2] and Girju, Badulescu and Moldovan in 2003 [6].

Berland and Charniak learn part-whole patterns from a part-of-speech tagged corpus, the Linguistic Data Consortium's (LDC) North American News Corpus (NANC). To illustrate the pattern learning phase they mention five example patterns. "whole's part", "part of {the|a} whole", "part in {the|a} whole", "parts of wholes", and "parts in wholes". The domain they used for evaluation is component/integral object relations between artifacts such as cars and windshields. Even though our domain is quite different, we found all five of their example patterns using our training data, respectively at rank 294, 290, 12, 128, and 2 (of 4502 learnt patterns).

Girju, Badulescu, and Moldovan, used the SemCor 1.7 corpus and the LA Times corpus from the Ninth Text Retrieval Conference (TREC-9). They used the meronyms from WordNet [9], mainly component/integral object and member/collection relations. Girju, Badulescu, and Moldovan also make the distinction between explicit and implicit part-whole constructions, but the implicit constructions they focus on are mainly possessive forms like "the girl's mouth", "eyes of the baby", "oxygen-rich water", and "high heel shoes". They list the three most frequent patterns, which also contain part-of-speech tags. "part of whole", "whole's part", and "part *Verb* whole". We found the first two patterns, as mentioned above, and many instances of the third pattern, such as "part fortified whole" at rank 4.

Other applications of part-whole relations than discovering sources of substances are query expansion for image retrieval [8, Ch. 6], and geographical retrieval [3].

8 Discussion

Our experimental setup assumes that all interesting information pertaining to some carcinogenic substance can be obtained in one single retrieval step. The construction of complex paths from the substance to the eventual exposure has to happen in the mind of the user—and depends solely on his expertise and ingenuity. This is a severe limitation

Acrylamide

concept (whole)	rank
coffee	18
fried food	22
plastics industry	39
smoke	42
drinking water	43
olives	103
paper	109
dyes	114
soil	144
fish	158
herbicide	181
water treatment	195
textiles	275
air	not found
baked food	not found

Benzene

concept (whole)	rank
leaded gasoline	1
water	4
solvents	9
smoke	10
dyes	32
pesticides	68
soil	69
detergents	76
cola	84[a]
rubber	161
bottled water	191
rivers	228
lubricants	340
air	not found[b]
fats	not found

[a] soft drinks appear at rank 5
[b] found air pollution and atmosphere

Asbestos

concept (whole)	rank
insulation	5
vermiculite	9
roofing	12
building materials	16
flooring	23
rocks	37
water	47
brakes	67
adhesives	127
cars	160
mucus	211
cement pipes	not found[a]
sewage	not found[b]
air	not found
feces	not found

[a] found cement and pipes
[b] found refuse and wastewater

Dioxins

concept (whole)	rank
fish	2[a]
paper	3
soil	7
herbicides	8
defoliants	17[b]
water	32
smoke	38
bleach	39
chickens	75
animal fat	106
animal feed	138
waste incineration	142
pigs	not found[c]
air	not found[d]
diesel trucks	not found[e]

[a] also found fishermen
[b] also found vietnam
[c] found cattle and livestock
[d] found air quality
[e] found exhaust gases

Table 5. Recall bases for four sample substances.

Prec.	D	pattern	Prec.	D	pattern
.84	26799	part in whole	.76	980	part content in the whole
.68	8787	whole with part	.96	745	part-treated whole
.84	5266	part in the whole	.84	786	part derived from whole
.96	4249	part from whole	.76	852	whole rich in part
.68	5917	part for whole	.28	2306	whole high part
.60	5794	part content whole	.88	617	part-containing whole
.88	3949	whole contain part	.20	2571	whole add part
1	2934	whole containing part	.72	700	part in most whole
.64	4415	part based whole	.80	623	part for use in whole
.72	3558	whole using part	.40	1169	part to make whole
.92	2591	part levels in whole	.72	630	add part to the whole
1	2336	part-laden whole	.72	580	part enriched whole
.84	2327	part content in whole	.56	703	part in many whole
1	1945	whole contains part	.96	404	part-enriched whole
.76	2536	whole have part	.72	527	part contents in whole
.72	2622	part into whole	.52	608	added part to whole
.88	2035	part is used in whole	.92	314	part occurs naturally in whole
1	1760	part found in whole	.84	288	part extracted from whole
.52	3217	part free whole	.96	226	whole enriched with part
1	1672	part is found in whole	.68	310	part to our whole
.88	1834	part-rich whole	.16	1160	whole provide part
.80	1994	part used in whole	.68	247	added part to the whole
.92	1680	part content of whole	.72	220	whole with added part
.20	7711	whole for part	.96	137	part found in many whole
.96	1497	part is present in whole	1	124	whole containing high part
.84	1600	add part to whole	.76	134	part replacement in whole
.88	1496	part added to whole	.60	133	part for making whole
.80	1597	part in their whole	.88	64	whole fortified with part
.92	1372	part-based whole	.76	74	whole have part added
.88	1421	part in these whole	.96	54	part-fortified whole
1	1218	whole that contain part	.36	120	part compound for whole
1	1203	part levels in the whole	.36	120	part fortified whole
.84	1361	part in all whole	1	24	whole sweetened with part
1	1112	part contained in whole	.16	89	whole preserves part
.76	1455	part in some whole	.91	11	part-reduced whole
.84	1301	part in your whole	.90	10	part gives the whole its
1	1058	part present in whole	.04	85	part sweetened whole
.76	1350	part in our whole	.27	11	part-increased whole
1	985	part laden whole	.67	3	part-added whole
.32	3052	whole use part	1	1	part-sweetened whole
.52	1648	whole mit part	1	1	part to sweeten their whole
.84	930	whole made with part	1	1	part fortification of whole
.88	885	part-free whole	0	0	part additions in various whole
.52	1477	part is in whole	0	0	part used in making whole
.80	945	part is added to whole	0	242	part hydrogenated whole
.92	811	whole high in part			

Table 6. The 91 patterns used for the learning of wholes, ordered by the number of correct pairs it yielded. Prec. is Precision approximated on a sample of 25 occurrences (or less if freq. < 25). D is the number of discovered part-whole phrases.

that leaves room for considerable improvement. A relatively straightforward extension would be to iterate the retrieval step using suitable wholes found in retrieval step $n-1$ in the part slot in retrieval step n. Separation of roles, classes, etc. amongst the wholes by means of classification (*cf., e.g.,* [7]) might be necessary to limit the inevitable loss of precision. For example, if step $n-1$ yielded that there is benzene in some fish, then proceeding to investigate in step n whether these fish are part of people's diet. If, however, step $n-1$ yielded that benzene is part of a group of carbon-based chemicals, then proceeding to investigate these chemicals might lead to excessive topic drift.

The usefulness of such an extension depends to a large extent on the validity of some sort of transitive reasoning over the paths. Yet, the transitivity characteristics of part-whole expressions are notoriously quirky. Existing accounts actually either take the classical route set out by Stanislaw Lesniewski in the 1920's, defining the relations in question axiomatically and with little consideration for actual usage, or they formulate reasoning patterns for specific application domains and expressions (*cf., e.g.,* [10]). Neither approach is applicable to the mixed bags of "interesting" token relations our setup derives from natural language usage. A rare attempt to ground reasoning patterns in the general usage of part-whole expressions is contained in [12]. Even though our lay-out is orthogonal (and not even coextensive) to their influential classification of part-whole relations, their basic intuition w.r.t. transitivity does carry over to our case. In short:

1. The part-whole relations, P, expressed in natural language form a partial order $\mathscr{P} = \langle P, \geq \rangle$;
2. The weakest link determines the interpretation of a chain of part-whole pairs w.r.t. transitivity;
3. Transitivity fails if the chain contains uncomparable relation instances (w.r.t. $\geq$).

Contrary to [12] we assume that there is some weakest mereological relation, i.e., the poset $\mathscr{P}$ has a minimum element. (2) can then be generalized as follows:

2' Any element of $\mathscr{P}$ which is compatible with (i.e., as least as weak as) every relation used to form a chain of part-whole pairs determines a transitive interpretation of that chain.

This means that for every chain of part-whole pairs there is a meaningful, albeit sometimes rather weak, transitive interpretation available. It depends solely on the intended utilization whether the information obtained in this way is specific enough to be useful. What has its merits in a task with a strong element of exploration and novelty detection like our use case, may well be a show stopper for tasks such as diagnosis in a process control environment. Refinements, especially concerning the classification of relation types and the properties of the poset of relations are necessary to extend the general applicability of this approach.

This is especially true when our work is placed in the more general context of vocabulary and ontology alignment. Most ontology-alignment systems aim at finding equivalence relations. Yet, many real-world alignment cases have to deal with vocabularies that have a different level of aggregation. (*cf.*, [11]) In such cases equivalent concepts are quite rare, while aggregation relations, such as broader/narrower term, subclass and part-whole, are common. The carcinogen-source discovery case can be seen as an

ontology-alignment problem where the alignment relation is the part-whole relation and the vocabularies are the controlled vocabulary of IARC group 1 carcinogens, and the AGROVOC and NALT thesauri. Under this perspective our work describes a first step towards a novel approach to ontology alignment. The influence part-whole alignment relations have on the consistency of the resulting aligned ontologies is unknown.

Acknowledgements

Margherita Sini and Johannes Keizer (FAO), Lori Finch (NAL), Fred van de Brug (TNO), Dean Allemang (BU), Alistair Miles (CCLRC) and Dan Brickley (W3C), the IARC, EPA, IFIC, and FDA, Vera Hollink (UvA), Sophia Katrenko (UvA), Mark van Assem (VUA), Laura Hollink (VUA), Véronique Malaisé (VUA). This work is part of the Virtual Lab e-Science project[14].

References

1. Alessandro Artale, Enrico Franconi, Nicola Guarino, and Luca Pazzi. Part-whole relations in object-centered systems: an overview. *Data & Knowledge Engineering*, 20(3):347–383, 1996.
2. Matthew Berland and Eugene Charniak. Finding parts in very large corpora. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999.
3. Davide Buscaldi, Paolo Rosso, and Emilio Sanchis Arnal. A wordnet-based query expansion method for geographical information retrieval. In *Working Notes for the CLEF 2005 Workshop*, 2005.
4. Oren Etzioni, Michael Cafarella, Doug Downey, Ana-Maria Popescu Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Methods for domain-independent information extraction from the web: An experimental comparison. In *Proc. of the AAAI Conference*, 2004.
5. Michal Finkelstein-Landau and Emmanuel Morin. Extracting semantic relationships between terms: Supervised vs. unsupervised methods. In *International Workshop on Ontological Engineering on the Global Information Infrastructure*, pages 71–80, 1999.
6. Roxana Girju, Adriana Badulescu, and Dan Moldovan. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proc. of the HLT-NAACL*, 2003.
7. Nicola Guarino and Christopher Welty. An overview of ontoclean. In Steffen Staab and Rudi Studer, editors, *Handbook on Ontologies*, pages 151–171. Springer Verlag, 2004.
8. Laura Hollink. *Semantic Annotation for Retrieval of Visual Resources*. PhD thesis, Free University Amsterdam, 2006. Submitted. URI: http://www.cs.vu.nl/∼laurah/thesis/thesis.pdf.
9. George A. Miller. Wordnet: a lexical database for english. *Communications of the ACM (CACM)*, 38(11):39–41, 1995.
10. Stefan Schulz and Udo Hahn. Part-whole representation and reasoning in formal biomedical ontologies. *Artificial Intelligence in Medicine*, 34(3):179–200, 2005.
11. Willem Robert van Hage, Sophia Katrenko, and Guus Schreiber. A method to combine linguistic ontology-mapping techniques. In *International Semantic Web Conference*, pages 732–744, 2005.
12. Morton E. Winston, Roger Chaffin, and Douglas Herrmann. A taxonomy of part-whole relations. *Cognitive Science*, 11:417–444, 1987.

[14] `http://www.vl-e.org`