On Learning Subtypes of the Part-Whole Relation: Do Not Mix your Seeds

Ashwin Ittoo
University of Groningen
Groningen, The Netherlands
r.a.ittoo@rug.nl

Gosse Bouma
University of Groningen
Groningen, The Netherlands
g.bouma@rug.nl

Abstract

An important relation in information extraction is the part-whole relation. Ontological studies mention several types of this relation. In this paper, we show that the traditional practice of initializing minimally-supervised algorithms with a single set that mixes seeds of different types fails to capture the wide variety of part-whole patterns and tuples. The results obtained with mixed seeds ultimately converge to one of the part-whole relation types. We also demonstrate that all the different types of part-whole relations can still be discovered, regardless of the type characterized by the initializing seeds. We performed our experiments with a state-of-the-art information extraction algorithm.

1 Introduction

A fundamental semantic relation in many disciplines such as linguistics, cognitive science, and conceptual modelling is the part-whole relation, which exists between parts and the wholes they compise (Winston et al., 1987; Gerstl and Pribbenow, 1995). Different types of part-whole relations, classified in various taxonomies, are mentioned in literature (Winston et al., 1987; Odell, 1994; Gerstl and Pribbenow, 1995; Keet and Artale, 2008). The taxonomy of Keet and Artale (2008), for instance, distinguishes part-whole relations based on their transitivity, and on the semantic classes of entities they sub-categorize. Part-whole relations are also crucial for many information extraction (IE) tasks (Girju et al., 2006). Annotated corpora and semantic dictionaries used in IE, such as the ACE corpus[1] and WordNet (Fellbaum, 1998), include examples of part-whole relations. Also, previous relation extraction work, such as Berland and Charniak (1999) and Girju et al. (2006), have specifically targeted the discovery of part-whole relations from text. Furthermore, part-whole relations are de-facto benchmarks for evaluating the performance of general relation extraction systems (Pantel and Pennacchiotti, 2006; Beamer et al., 2008; Pyysalo et al., 2009). However, these relation extraction efforts have overlooked the ontological distinctions between the different types of part-whole relations. They assume the existence of a single relation, subsuming the different part-whole relation types.

In this paper, we show that enforcing the ontological distinctions between the different types of part-whole relations enable information extraction systems to capture a wider variety of both generic and specialised part-whole lexico-syntactic patterns and tuples. Specifically, we address 3 major questions.

1. Is information extraction (IE) harder when learning the individual types of part-whole relations? That is, we determine whether the performance of state-of-the-art IE systems in learning the individual part-whole relation types increases (due to more coherency in the relations' linguistic realizations) or drops (due to fewer examples), compared to the traditional practice of considering a single part-whole relation.

2. Are the patterns and tuples discovered when focusing on a specific part-whole relation type confined to that particular type? That is, we investgate whether IE systems discover examples representative of the different types by targetting one particular part-whole relation type.

3. Are more distinct examples discovered when IE systems learn the individual part-whole relation types? That is, we determine whether

[1] http://projects.ldc.upenn.edu/ace/

a wider variety of unique patterns and tuples are extracted when IE systems target the different types of part-whole relations instead of considering a single part-whole relation that subsumes all the different types.

To answer these questions, we bootstrapped a minimally-supervised relation extraction algorithm, based on Espresso (Pantel and Pennacchiotti, 2006), with different seed-sets for the various types of part-whole relations, and analyzed the harvested tuples and patterns.

2 Previous Work

Investigations on the part-whole relations span across many disciplines, such as conceptual modeling (Artale et al., 1996; Keet, 2006; Keet and Artale, 2008), which focus on the ontological aspects, and linguistics and cognitive sciences, which focus on natural language semantics. Several linguistically-motivated taxonomies (Odell, 1994; Gerstl and Pribbenow, 1995), based on the work of Winston et al. (1987), have been proposed to clarify the semantics of the different part-whole relations types across these various disciplines. Keet and Artale (2008) developed a formal taxonomy, distinguishing transitive *mereological* part-whole relations from intransitive *meronymic* ones. Meronymic relations identified are: 1) *member-of*, between a physical object (or role) and an aggregation, e.g. *player-team*, 2) *constituted-of*, between a physical object and an amount of matter e.g. *clay-statue*, 3) *sub-quantity-of*, between amounts of matter or units, e.g. *oxygen-water* or *m-km*, and 4)*participates-in*, between an entity and a process e.g. *enzyme-reaction*. Mereological relations are: 1)*involved-in*, between a phase and a process, e.g. *chewing-eating*, 2) *located-in*, between an entity and its 2-dimensional region, e.g. *city-region*, 3)*contained-in*, between an entity and its 3-dimensional region, e.g.*tool-trunk*, and 4)*structural part-of*, between integrals and their (functional) components, e.g. *engine-car*. This taxonomy further discriminates between part-whole relation types by enforcing semantical selectional restrictions, in the form of DOLCE ontology (Gangemi et al., 2002) classes, on their entities.

In NLP, information extraction (IE) techniques, for discovering part-whole relations from text have also been developed. Berland and Charniak (1999) use manually-crafted patterns, similar to Hearst (1992), and on initial "seeds" denoting "whole" objects (e.g. building) to harvest possible "part" objects (e.g. room) from the North Americal News Corpus (NANC) of 1 million words. They rank their results with measures like log-likelihood (Dunning, 1993), and report a maximum accuracy of 70% over their top-20 results. In the supervised approaches in Girju et al. (2003) and Girju et al. (2006), lexical patterns expressing part-whole relations between WordNet concept pairs are manually extracted from 20,000 sentences of the L.A Times and SemCor corpora (Miller et al., 1993), and used to generate a training corpus, with manually-annotated positive and negative examples of part-whole relations. Classification rules, induced over the training data, achieve a precision of 80.95% and recall of 75.91% in predicting whether an unseen pattern encode a part-whole relation. Van Hage et al. (2006) acquire 503 part-whole pairs from dedicated thesauri (e.g. AGROVOC[2]) to learn 91 reliable part-whole patterns. They substituted the patterns' "part" arguments with known entities to formulate web-search queries. Corresponding "whole" entities were then discovered from documents in the query results with a precision of 74%. The part-whole relation is also a benchmark to evaluate the performance of general information extraction systems. The Espresso algorithm (Pantel and Pennacchiotti, 2006) achieves a precision of 80% in learning part-whole relations from the Acquaint (TREC-9) corpus of nearly 6M words. Despite the reasonable performance of the above IE systems in discovering part-whole relations, they overlook the ontological distinctions between the different relation types. For example, Girju et al. (2003) and Girju et al. (2006) assume a single part-whole relation, encompassing all the different types mentioned in the taxonomy of Winston et al. (1987). Similarly, the minimally-supervised Espresso algorithm (Pantel and Pennacchiotti, 2006) is initialized with a single set that mixes seeds of heterogeneous types, such as *leader-panel* and *oxygen-water*, which respectively correspond to the *member-of* and *sub-quantity-of* relations in the taxonomy of Keet and Artale (2008).

[2] http://aims.fao.org/website/AGROVOC-Thesaurus/sub

3 Methodology

Our aim is to compare the relations harvested when a minimally-supervised IE algorithm is initialized with separate sets of seeds for each type of part-whole relation, and when it is initialized following the traditional practice of a single set that mixes seeds of the different types. To distinguish between types of part-whole relations, we commit to the taxonomy of Keet and Artale (2008) (Keet's taxonomy), which uses sound ontological formalisms to unambiguously discrimate the relation types. Also, this taxonomy classifies the various part-whole relations introduced in literature, including ontologically-motivated mereological relations and linguistically-motivated meronymic ones. We adopt a 3-step approach to address our questions from section 1.

1. Define prototypical seeds (part-whole tuples) as follows:
   - (Separate) sets of seeds for each type of part-whole relation in Keet's taxonomy.
   - A single set that mixes seeds denoting all the different part-whole relations types.
2. Part-whole relations extraction from a corpus by initializing a minimally-supervised IE algorithm with the seed-sets
3. Evaluation of the harvested relations to determine performance gain/loss, types of part-whole relations extracted, and distinct and unique patterns and tuples discovered.

The corpora and IE algorithm we used, and the seed-sets construction are described below. Results are presented in the next section.

3.1 Corpora

We used the English and Dutch Wikipedia texts since their broad-coverage and size ensures that they include sufficient lexical realizations of the different types of part-whole relations. Wikipedia has also been targeted by recent IE efforts (Nguyen et al., 2007; Wu and Weld, 2007). However, while they exploited the structured features (e.g. infoboxes), we only consider the unstructured texts. The English corpus size is approximately 470M words ($\sim$ 80% of the August 2007 dump), while for Dutch, we use the full text collection (February 2008 dump) of approximately 110M words. We parsed the English and Dutch corpora respectively with the Stanford[3] (Klein and Manning, 2003) and the Alpino[4] (van Noord, 2006) parsers, and formalized the relations between terms (entities) as dependency paths. A dependency path is the shortest path of lexico-syntactic elements, i.e. shortest lexico-syntactic pattern, connecting entities (proper and common nouns) in their parse-trees. Such a formalization has been successfully employed in previous IE tasks (see Stevenson and Greenwood (2009) for an overview). Compared to traditional surface-pattern representations, used by Pantel and Pennacchiotti (2006), dependency paths abstract from surface texts to capture long range dependencies between terms. They also alleviate the manual authoring of large numbers of surface patterns. In our formalization, we substitute entities in the dependency paths with generic placeholders PART and WHOLE. Below, we show two dependency paths (1-b) and (2-b), respectively derived from English and Dutch Wikipedia sentences (1-a) and (2-a), and denoting the relations between *sample-song*, and *alkaloïde-plant*.

(1) a. The song "Mao Tse Tung Said" by Alabama 3 contains samples of a speech by Jim Jones
    b. WHOLE+nsubj $\leftarrow$ contains $\rightarrow$ dobj+PART

(2) a. Alle delen van de planten bevatten alkaloïden en zijn daarmee giftig (*All parts of the plants contain alkaloids and therefore are poisonous*)
    b. WHOLE+obj1+van+mod+deel+su $\leftarrow$ bevat$\rightarrow$ obj1+PART

In our experiments, we only consider those entity pairs (tuples), patterns, and co-occuring pairs-patterns with a minimum frequency of 10 in the English corpus, and 5 in the Dutch corpus. Statistics on the number of tuples and patterns preserved after applying the frequency cut-off are given in Table 1.

3.2 Information Extraction Algorithm

As IE algorithm for extracting part-whole relations from our texts, we relied on *Espresso*, a minimally-supervised algorithm, as described by Pantel and Pennacchiotti (2006). They show

[3] http://nlp.stanford.edu/software/lex-parser.shtml

[4] http://www.let.rug.nl/~vannoord/alp/Alpino

	English	Dutch
words	470.0	110.0
pairs	328.0	28.8
unique pairs	6.7	1.4
patterns	238.0	54.0
unique patterns	2.0	0.9

Table 1: Corpus Statistics in millions

that the algorithm achieves state-of-the-art performance when initialized with relatively small seed-sets over the Acquaint corpus ($\sim$ 6M words). Recall is improved with web search queries as additional source of information.

Espresso extracts surface patterns connecting the seeds (tuples) in a corpus. The reliability of a pattern p, $r(p)$, given a set of input tuples I, is computed using (3), as its average strength of association with each tuple,i, weighted by each tuple's reliability, $r_\iota(i)$.

$$r_\pi(p) = \frac{\sum_{i \in I} \left(\frac{pmi(i,p)}{\max_{pmi}} \times r_\iota(i) \right)}{|I|} \quad (3)$$

In this equation, $pmi(i, p)$ is the pointwise mutual information score (Church and Hanks, 1990) between a pattern, p (e.g. *consist-of*), and a tuple, i (e.g. *engine-car*), and max_{pmi} is the maximum PMI score between all patterns and tuples. The reliability of the initializing seeds is set to 1.

The top-k most reliable patterns are selected to find new tuples. The reliability of each tuple i, $r_\iota(i)$ is computed according to (4), where P is the set of harvested patterns. The top-m most reliable tuples are used to infer new patterns.

$$r_\iota(i) = \frac{\sum_{i \in I} \left(\frac{pmi(i,p)}{\max_{pmi}} \times r_\pi(p) \right)}{|P|} \quad (4)$$

The recursive discovery of patterns from tuples and vice-versa is repeated until a threshold number of patterns and/or tuples have been extracted. In our implementation, we maintain the core of the original Espresso algorithm, which pertains to estimating the reliability of patterns and tuples.

Pantel and Pennacchiotti (2006) mention that their method is independent of the way patterns are formulated. Thus, instead of relying on surface patterns, we use dependency paths (as described above). Another difference is that while Pantel and Pennacchiotti (2006) complement their small corpus with documents retrieved from the web, we only rely on patterns extracted from our (much larger) corpora. Finally, we did not apply the discounting factor suggested in Pantel and Pennacchiotti (2006) to correct for the fact that PMI overestimates the importance of low-frequency events. Instead, as explained above, we applied a general frequency cut-off.[5]

3.3 Seed Selection

Initially,we selected seeds from WordNet (Fellbaum, 1998) (for English) and EuroWordNet (Vossen, 1998) (for Dutch) to initialize the IE algorithm. However, we found that these pairs, such as *acinos-mother of thyme* or *radarscherm-radarapparatuur (radar screen - radar equipment*, hardly co-occured with reasonable frequency in Wikipedia sentences, hindering pattern extraction. We therefore adopted the following strategy.

We searched our corpora for archetypal patterns, e.g. *contain* , which characterize all the different types of part-whole relations. The tuples sub-categorized by these patterns in the English texts were automatically[6] typed to appropriate DOLCE ontology[7] classes, corresponding to those employed by Keet and Artale for constraining the entity pairs participating in different types of part-whole relations. The types of part-whole relations instantiated by the tuples could then be determined based on their ontological classes. Separate sets of 20 tuples, with each set corresponding to a specific relation type in the taxonomy of Keet and Artale (Keet's taxonomy), were then created. For example, the English Wikipedia tuple $t1$ =*actor-cast* was used as a seed to discover *member-of* part-whole relations since both its elements were typed to the SOCIAL OBJECT class of the DOLCE ontology, and according to Keet's taxonomy, they instantiate a *member-of* relation. Seeds for extracting relations from the Dutch corpus were defined in a similar way, except that we manually determined their ontological classes based on the class glossary of DOLCE.

Below, we only report on the *member-of* and *sub-quantity-of* meronymic relations, and on the *located-in*, *contained-in* and *structural part-of* mereological relations. We were unable to find sufficient seeds for the *constituted-of* meronymic

[5] We experimented with the suggested discounting factor for PMI, but were not able to improve over the accuracy scores reported later.

[6] Using the Java-OWL API, from `http://protege.stanford.edu/plugins/owl/api/`

[7] OWL Version 0.72, downloaded from `http://www.loa-cnr.it/DOLCE.html/`

Lg	Part	Whole	#	Type
EN	grave	church	155	contain
NL	beeld	kerk	120	contain
	(*statue*)	(*church*)		
EN	city	region	3735	located
NL	abdij	gemeente	36	located
	(*abbey*)	(*community*)		
EN	actor	cast	432	member
NL	club	voetbal_bond	178	member
	(*club*)	(*soccer union*)		
EN	engine	car	3509	structural
NL	geheugen	computer	14	structural
	(*memory*)	(*computer*)		
EN	alcohol	wine	260	subquant
NL	alcohol	bier	28	subquant
	(*alcohol*)	(*beer*)		

Table 2: Seeds used for learning part-whole relations (*contained-in, located-in, member-of, structural part-of, sub-quantity-of*).

relations (e.g. *clay-statue*). Also, we did not experiment with the *participates-in* and *involved-in* relations since their lexical realizations in our corpora are sparse, and they contain at least one verbal argument, whereas we only targeted patterns connecting nomimals. Sample seeds, their corpus frequency, and the part-whole relation type they instantiate from the English (EN) and Dutch (NL) corpora are illustrated in Table 2. Besides the five specialized seed-sets of 20 prototypical tuples for the aforementioned relations, we also defined a *general* set of mixed seeds, which combines four seeds from each of the specialized sets.

4 Experiments and Evaluation

We initialized our IE algorithm with the seed-sets to extract part-whole relations from our corpora. The same parameters as Pantel and Pennacchiotti (2006) were used. That is, the 10 most reliable patterns inferred from the initial seeds are bootstrapped to induce 100 part-whole tuples. In each subsequent iteration, we learn one additional pattern and 100 additional tuples. We evaluated our results after 5 iterations since the performance in later iterations was almost constant. The results are discussed next.

	meronomic		mereological			
	memb	subq	cont	struc	locat	gen
EN	0.67	0.74	0.70	0.82	0.75	0.80
NL	0.68	0.60	0.60	0.60	0.70	0.71

Table 3: Precision for seed-sets representing specific types of part-whole relations (*member-of, sub-quantity-of, contained-in, structural part-of* and *located-in*), and for the *general* set composed of all types.

4.1 Precision of Extracted Relations

Two human judges manually evaluated the tuples extracted from the English and Dutch corpora per seed-set in each iteration of our algorithm. Tuples that unambiguously instantiated part-whole relations were considered *true positives*. Those that did not were considered *false positives*. Ambiguous tuples were discarded. The precision of the tuples discovered by the different seed-sets in the last iteration of our algorithm are in Table 3.

These results reveal that the precision of harvested tuples varies depending on the part-whole relation type that the initializing seeds denote. Mereological seeds (*cont, struct, locat* sets) outperformed their meronymic counterparts (*memb, subq*) in extracting relations with higher precision from the English texts. This could be attributed to their formal ontological grounding, making them less ambiguous than the linguistically-motivated meronymic relations (Keet, 2006; Keet and Artale, 2008). The precision variations were less discernible for tuples extracted from the Dutch corpus, although the best precision was still achieved with mereological *located-in* seeds. We also noticed that the precision of tuples extracted from both the English and Dutch corpora by the general set of mixed seeds was as high as the maximum precision obtained by the individual sets of specialized seeds over these two corpora, i.e. 0.80 (*general* seeds) vs. 0.82 (*structural part-of* seeds) for English, and 0.71 (*general* seeds) vs. 0.70 (*located-in* seeds) for Dutch. Based on these findings, we address our first question, and conclude that 1) the type of relation instantiated by the initializing seeds affects the performance of IE algorithms, with mereological seeds being in general more fertile than their meronymic counterparts, and generating higher-precision tuples; 2) the precision achieved when initializing IE algorithms with a general set, which mixes

seeds of heterogeneous part-whole relation types, is comparable to the best results obtained with individual sets of specialized seeds, denoting specific part-whole relations. An evaluation of the patterns and tuples extracted indicated considerable precision drop between successive iterations of our algorithm. This appears to be due to *semantic drift* (McIntosh and Curran, 2009), where highly-ambiguous patterns promote incorrect tuples , which in turn, compound the precision loss.

4.2 Types of Extracted Relations

Initializing our algorithm with seeds of a particular type always led to the discovery of tuples characterizing other types of part-whole relations in the English corpus. This can be explained by prototypical patterns, e.g. "include", generated regardless of the seeds' types, and which are highy correlated with, and hence, trigger tuples denoting other part-whole relation types. An almost similar observation was made for the Dutch corpus, except that tuples instantiating the *member-of* relation could only be learnt using initial seeds of that particular type (i.e. *member-of*). Upon inspecting our results, it was found that this phenomenon was due to the distinct and specific patterns, such as "treedt toe tot" ("become member of"), which linguistically realize the *member-of* relations in the Dutch corpus. Thus, initializing our IE algorithm with seeds that instantiate relations other than *member-of* fails to detect these unique patterns, and fails to subsequently discover part-whole tuples describing the *member-of* relations. Our findings are illustrated in Table 4, where each cell lists a tuple of a particular type (column), which was harvested from seeds of a given type (row). These results answer our second question.

4.3 Distinct Patterns and Tuples

We address our third question by comparing the output of our algorithm to determine whether the results obtained by initializing with the individual specialized seeds were (dis)similar and/or distinct. Each result set consisted of maximally 520 tuples (including 20 initializing seeds) and 15 lexico-syntactic patterns, obtained after five iterations.

Tuples extracted from the English corpus using the *member-of* and *contained-in* seed-sets exhibited a high degree of similarity, with 465 common tuples discovered by both sets. These identical tuples were also assigned the same ranks (reliability) in the results generated by the *member-of* and *contained-in* seeds, with a Spearman rank correlation of 0.82 between their respective outputs. This convergence was also reflected in the fact that the *member-of* and *contained-in* seeds generated around 80% of common patterns. These patterns were mostly prototypical ones indicative of part-whole relations, such as WHOLE+nsubj $\leftarrow$ include $\rightarrow$ dobj+PART ("include") and their cognates involving passive forms and relative clauses. However, the specialized seeds also generated distinct patterns, like "joined as" and "released with" for the *member-of* and *contained-in* seeds respectively.

The most distinct tuples and patterns were harvested with the *sub-quantity-of*, *structural part-of*, and *located-in* seeds. Negative Spearman correlation scores were obtained when comparing the results of these three sets among themselves, and with the results of the *member-of* and *contained-in* seeds, indicating insignificant similarity and overlap. Examining the patterns harvested by the *sub-quantity-of*, *structural part-of*, and *located-in* seeds revealed a high prominence of specialized and unique patterns, which specifically characterize these relations. Examples of such patterns include "made with", "released with" and "found in", which lexically realize the *sub-quantity-of*, *structural part-of*, and *located-in* relations respectively.

For the Dutch corpus, the seeds that generated the most similar tuples were those corresponding to the *sub-quantity-of*, *contained-in*, and *structural part-of* relations, with 490 common tuples discovered, and a Spearman rank correlation in the range of 0.89-0.93 between their respective outputs. As expected, these seeds also led to the discovery of a substantial number of common and prototypical part-whole patterns. Examples include "bevat" ("contain"), "omvat" ("comprise"), and their variants. The most distinct results were harvested by the *located-in* and *member-of* seeds, with negative Spearman correlation scores between the output tuples indicating hardly any overlap. We also found out that the patterns harvested by the *located-in* and *member-of* seeds characteristically pertained to these relations. Example of such patterns include "ligt in" ("lie in"), "is gelegen in" ("is located in"), and "treedt toe tot" ("become member of"), respectively describing the *located-in* and *member-of* relations.

Thus, we observed that 1) tuples harvested from

	Tuples→ *Seeds*↓	meronomic member	meronomic subquant	mereological contained	mereological struct	mereological located
EN	member	ship-convoy	alcohol-wine	card-deck	proton-nucleus	lake-park
	subquant	aircraft-fleet	moisture-soil	building-complex	engine-car	commune-canton
	contained	aircraft-fleet	alcohol-wine	relic-church	base-spacecraft	campus-city
	structural	brother-family	mineral-bone	library-building	inlay-fingerboard	hamlet-town
	located	performer-cast	alcohol-blood	artifact-museum	chassis-car	city-shore
NL	member	sporter-ploeg (*athlete-team*)	helium-atmosfeer (*helium-atmosphere*)	stalagmieten-grot (*stalagnites-cave*)	shirt-tenue (*shirt-outfit*)	boerderij-dorp (*farm-village*)
	subquant	—	vet-kaas (*fat-cheese*)	pijp_orgel-kerk (*pipe-organ-church*)	kam-gitaar (*bridge-guitar*)	paleis-stad (*palace-city*)
	contained	—	tannine-wijn (*tannine-wine*)	kamer-toren (*room-tower*)	atoom-molecule (*atom-molecule*)	paleis-stad (*palace-city*)
	structural	—	kinine-tonic (*quinine-tonic*)	beeld-kerk *statue-church*)	wervel-ruggengraat (*vertebra-backbone*)	paleis-stad (*palace-city*)
	located	—	—	kunst_werk-kathedraal (*work of art-cathedral*)	poort-muur (*gate-wall*)	metro_station-wijk (*metro station-quarter*)

Table 4: Sample tuples found per relation type.

both the English and Dutch corpora by seeds instantiating a single particular type of part-whole relation highly correlated with tuples discovered by at least one other type of seeds (*member-of* and *contained-in* for English, and *sub-quantity-of*, *contained-in* and *structural part-of* for Dutch); 2) some part-whole relations are manifested by a wide variety of specialized patterns (*sub-quantity-of*, *structural part-of*, and *located-in* for English, and *located-in* and *member-of* for Dutch).

Finally, instead of a single set that mixes seeds of different types, we created five such *general* sets by picking four different seeds from each of the specialized sets, and used them to initialize our algorithm. When examining the results of each of the five *general* sets, we found out that they were unstable, and always correlated with the output of a different specialized set.

Based on these findings, we believe that the traditional practice of initializing IE algorithms with *general* sets that mix seeds denoting different part-whole relation types leads to inherently unstable results. As we have shown, the relations extracted by combining seeds of heterogeneous types almost always converge to one specific part-whole relation type, which cannot be conclusively predicted. Furthermore, *general* seeds are unable to capture the specific and distinct patterns that lexically realize the individual types of part-whole relations.

5 Conclusions

In this paper, we have investigated the effect of ontologically-motivated distinctions in part-whole relations on IE systems that learn instances of these relations from text.

We have shown that learning from specialized seeds-sets, denoting specific types of the part-whole relations, results in precision that is as high as or higher than the precision achieved with a general set that mixes seeds of different types. By comparing the outputs generated by different seed-sets, we observed that the tuples learnt with seeds denoting a specific part-whole relation type are not confined to that particular type. In most case, we are still able to discover tuples across all the different types of part-whole relations, regardless of the type instantiated by the initializing seeds. Most importantly, we demonstrated that IE algorithms initialized with general sets of mixed seeds harvest results that tend to converge towards a specific type of part-whole relation. Conversely, when starting with seeds representing a specific type, it is likely to discover tuples and patterns that are completely distinct from those found by a mixed seed-set.

Our results also illustrate that the outputs of IE algorithms are heavily influenced by the initializing seeds, concurring with the findings of McIntosh and Curran (2009). We believe that our results show a drastic form of this phenomenon: given a set of mixed seeds, denoting heterogeneous relations, the harvested tuples may converge towards any of the relations instantiated by the seeds. Predicting the convergent relation is in usual cases impossible, and may depend on factors pertaining to corpus characteristics. This instability strongly suggests that seeds instantiating different types of relations should not be mixed, partic-

ularly when learning part-whole relations, which are characterized by many subtypes. Seeds should be defined such that they represent an ontologically well-defined class, for which one may hope to find a coherent set of extraction patterns.

Acknowledgement

Ashwin Ittoo is part of the project "Merging of Incoherent Field Feedback Data into Prioritized Design Information (DataFusion)" (`http://www.iopdatafusion.org//`), sponsored by the Dutch Ministry of Economic Affairs under the IOP-IPCR program.

Gosse Bouma acknowledges support from the Stevin LASSY project (`www.let.rug.nl/~vannoord/Lassy/`).

References

A. Artale, E. Franconi, N. Guarino, and L. Pazzi. 1996. Part-whole relations in object-centered systems: An overview. *Data & Knowledge Engineering*, 20(3):347–383.

B. Beamer, A. Rozovskaya, and R. Girju. 2008. Automatic semantic relation extraction with multiple boundary generation. In *Proceedings of the 23rd national conference on Artificial intelligence-Volume 2*, pages 824–829. AAAI Press.

Matthew Berland and Eugene Charniak. 1999. Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64, Morristown, NJ, USA. Association for Computational Linguistics.

K.W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):74.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT, Cambridge.

A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. 2002. Sweetening ontologies with DOLCE. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web, Lecture Notes in Computer Science*, pages 223–233.

P. Gerstl and S. Pribbenow. 1995. Midwinters, end games, and body parts: a classification of part-whole relations. *International Journal of Human Computer Studies*, 43:865–890.

R. Girju, A. Badulescu, and D. Moldovan. 2003. Learning semantic constraints for the automatic discovery of part-whole relations. In *Proceedings of HLT/NAACL*, volume 3, pages 80–87.

R. Girju, A. Badulescu, and D. Moldovan. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.

M.A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics-Volume 2*, pages 539–545. Association for Computational Linguistics Morristown, NJ, USA.

C.M. Keet and A. Artale. 2008. Representing and reasoning over a taxonomy of part–whole relations. *Applied Ontology*, 3(1):91–110.

C.M. Keet. 2006. Part-whole relations in object-role models. *On the Move to Meaningful Internet Systems 2006, Lecture Notes in Computer Science*, 4278:1118–1127.

D. Klein and C.D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics Morristown, NJ, USA.

T. McIntosh and J.R. Curran. 2009. Reducing semantic drift with bagging and distributional similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Confe rence on Natural Language Processing of the AFNLP*, pages 396–404.

G.A. Miller, C. Leacock, R. Tengi, and R.T. Bunker. 1993. A semantic concordance. In *Proceedings of the 3rd DARPA workshop on Human Language Technology*, pages 303–308. New Jersey.

D.P.T. Nguyen, Y. Matsuo, and M. Ishizuka. 2007. Relation extraction from wikipedia using subtree mining. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 1414. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

J. Odell. 1994. Six different kinds of composition. *Journal of Object-Oriented Programming*, 5(8):10–15.

Patrick Pantel and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06)*, pages 113–120, Sydney, Australia.

S. Pyysalo, T. Ohta, J.D. Kim, and J. Tsujii. 2009. Static relations: a piece in the biomedical information extraction puzzle. In *Proceedings of the Workshop on BioNLP*, pages 1–9. Association for Computational Linguistics.

Mark Stevenson and Mark Greenwood. 2009. Dependency pattern models for information extraction. *Research on Language and Computation*, 3:13–39.

W.R. Van Hage, H. Kolb, and G. Schreiber. 2006. A method for learning part-whole relations. *The Semantic Web - ISWC 2006, Lecture Notes in Computer Science*, 4273:723–735.

Gertjan van Noord. 2006. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42. Presses univ. de Louvain.

P. Vossen, editor. 1998. *EuroWordNet A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic publishers.

M.E. Winston, R. Chaffin, and D. Herrmann. 1987. A taxonomy of part-whole relations. *Cognitive science*, 11(4):417–444.

F. Wu and D.S. Weld. 2007. Autonomously semantifying wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM.