Semi-automatic Induction of Underspecified Semantic Classes

Noriko Tomuro
DePaul University
School of Computer Science, Telecommunications and Information Systems
243 S. Wabash Ave.
Chicago IL 60604
cphdnt@ted.cs.depaul.edu

Abstract

This paper describes a semi-automatic method of inducing *underspecified semantic classes* from WordNet verbs and nouns. An underspecified semantic class is an abstract semantic class which represents *systematic polysemy*: a set of word senses that are related in systematic and predictable ways. The method first applies a simple statistical technique to extract sense cooccurrence from the ambiguous words, and creates a *type dependency graph* after a manual filtering step. Then, the underspecified classes are automatically induced by partitioning the ambiguous senses according to the nodes covered in the graph. We discuss the advantages and difficulties of our method by comparing our results and those in CORELEX (Buitelaar, 1997, 1998).

1 Introduction

Representation of *polysemy* has been one of the central issues in lexical semantics. The key problem is how to encode the multiple word senses which fluctuate depending on the particular context in which the word is used. For example, a noun **chicken** has (at least) two meanings, ANIMAL and MEAT, either of which can be determined in a given discourse context. A naive approach to representing polysemy is to enumerate all senses in a word entry, as is often seen in conventional dictionaries or machine readable dictionaries (MRDs). In fact, recent word sense disambiguation (WSD) research is based on sense enumeration as the fundamental lexical representation model.[1] However, as it has been pointed out in many previous literature, sense enumeration approach has severe limitations. First, it is impossible to statically define all possible, seemingly infinite number of word senses that arise in different contexts. Second, same or very similar senses of different words (eg. synonyms) are represented separately under each of these words, therefore the lexicon contains much redundancy. Third, sense enumeration does not encode any associations between the senses of a word, therefore it ignores the notion of relatedness or similarity of word senses. Consequently, sense enumeration does not support any linguistic generalization at all.

To overcome those difficulties, many approaches have been proposed which capture the regularities among polysemous words and explicitly represent the inter-relations between word senses. In particular, there have been several approaches which base the generalization on *systematic polysemy* (or *regular polysemy* (Apresjan, 1974)):[2] a set of word senses that are related

[1]See (Ng and Zelle, 1997) for a survey of current WSD techniques.
[2]Note that systematic polysemy should be contrasted with *homonymy* which refers to words which have more

in systematic and predictable ways. For example, **chicken** is systematically polysemous because MEAT meaning is predictable from ANIMAL since chicken as meat refers to the flesh of a chicken as a bird that is used for food. In those approaches that are based on systematic polysemy, systematic relations between word senses are captured in various ways, for instance by *lexical rules* (Ostler and Atkins, 1992; Copestake and Briscoe, 1992, 1995) expressed by if-then implication rules, or by (abstract) semantic types and the associated type shifting mechanisms (in *Generative Lexicon Theory (GL)* (Pusteovsky, 1995)), or by default inheritance in a hierarchically organized lexicon (Kilgarriff, 1992; Copestake and Briscoe, 1992, 1995; Verspoor, 1997). What's common in those approaches is that the linguistic generalization based on systematic polysemy can facilitate the notion of sense derivation: one sense is derivable from the other sense(s) of the word, either procedurally (in lexical rules and GL) or declaritively (in default inheritance). Then, the consequence is the lexical representation which is less redundant and has much more inferential properties.

Based on *GL*, (Buitelaar, 1997, 1998) proposes CORELEX, a broad-coverage database of *underspecified semantic classes* as a general ontology for systematic polysemy, and derives them from WordNet (Miller, 1990). An underspecified semantic class is an abstract semantic type/label that signifies a relation between systematically polysemous senses. From fine-grained word senses enumerated in WordNet, a set of coarse-grained basic types are selected, and then certain combinations of them that are in a systematic relation are grouped together under an abstract semantic class. This abstract class is underspecified because it does not specify either one of the member types. Thus, this approach relates the systematicity of polysemy to underspecified classes through the composition of basic types. In this underspecified approach, polysemous senses/types are treated collectively as a multi-dimensional semantic entity that encompasses several semantic "axes". Then, such entities (i.e., the underspecified classes) can be refined in a given context (analogous to the traditional notion of sense disambiguation), or receive a simultaneous interpretation of the member types. In the latter case, underspecified classes facilitate abductive inference in the context: given a word with n related senses, the identification of one sense in a context can imply maximally all n senses, some of which may only be implicit in the context. Then, the underspecified classes which cover broad domains can be used as initial semantic labels in the lexical representations which support deep contextual reasoning, or as general *background* lexicon for domain-specific reasoning tasks such as Information Extraction (Kilgarriff, 1997), as well as for many other natural language processing tasks that require contextual inferences.

However, automatic acquisition of systematic polysemy has been a difficult task. In fact, in most previous work in lexical semantics it is done manually (Buitelaar, 1997, 1998). In this paper, we present a semi-automatic method of inducing underspecified semantic classes from WordNet verbs and nouns. The method first applies a simple statistical technique to extract sense cooccurrence from the ambiguous words, and creates a *type dependency graph* after a manual filtering step. Then, the underspecified classes are automatically induced by partitioning the ambiguous senses according to the nodes covered in the graph. Although the method is not fully automated, it provides a principled way of acquiring systematic polysemy from a large-scale lexical resource, and reduces the amount of manual effort that was previously required. Furthermore, by having a manual intervention, the results will be able to reflect our prior knowledge about WordNet that was not assumed in the statistical analysis.

---

than one unrelated sense (eg. FINANCIAL_INSTITUTION and SLOPING_LAND meanings of the word **bank**)

2 Systematic Polysemy

Before presenting the induction method, we first clarify what we consider a systematic polysemy in the work described in this paper, and explain the assumptions we made for (semi-)automatically acquiring systematic polysemies.

2.1 Systematic Relations

Our systematic polysemy is analogous to *logical polysemy* in (Pusteovsky, 1995): word senses in which there is no change in lexical category, and the multiple senses of the word have overlapping, dependent, or shared meanings. This definition excludes meanings obtained by cross-categorical alternations (eg. denominals) or morphological alternations (eg. suffixing with *-ify*), or homonyms or metaphors, and includes only the senses of a word of the same category and literal form that have some systematic relations.

The definition of what we considered a systematic relation is the same with that in (Buitelaar, 1997, 1998), where the relation between two types are represented by an open and closed *dot* ($\circ$ for open and $\bullet$ for closed). The use of these dots are motivated by discourse reference: for types τ_1 and τ_2, $\tau_1 \circ \tau_2$ is used when the two types are never interpreted simultaneously in one discourse, whereas $\tau_1 \bullet \tau_2$ is used when the two types are always interpreted simultaneously. Therefore, senses in the open dot relation often denote different objects/actions, whereas senses in the closed dot relation denote different aspects of the same object/action. For example, ANIMAL and MEAT meanings of the noun **chicken** refer to two distinct objects, and they are usually not used in the same context (thus ANIMAL $\circ$ MEAT). On the other hand, CREATION and COMMUNICATION meanings of the verb **write** in a sentence "John wrote the book" refer to the same action of John writing the particular book, but CREATION describes the *causation* aspect, whereas COMMUNICATION describes the *purpose* aspect of the **write** action.[3]

2.2 Computational Assumptions

In order to compute the systematic relations and automatically acquire from WordNet, we had to make further assumptions. As with CORELEX, our underspecified semantic class is a combination of abstract semantic types (which we will call *basic types*). But, instead of manually determining whether a given set of basic types constitute a systematic relation as is done in (Buitelaar, 1997, 1998), we construct an underspecified class automatically through the composition of binary relations between the basic types.

However, there is one problem with this scheme, namely, a composition of independent (binary) relations may result in an implausible relation when they are taken together in some cases. Such an implausible composition can be detected, if it was to be done manually, by hypothesizing the decomposition of the member types and some contexts in which the interpretation is possible. Unfortunately it is difficult to do this processing automatically. In order to deal with this problem, we brought in the notion of sense *directionality* to control implausible compositions. We first define a relation between two basic types to be directed, from the primary type to the secondary type, and then allow the compositions (or inferences) to be made only in one direction. The directionality of a relation here is essentially a dependency relation, in which the primary type cause the existence of the secondary type(s).

[3] In this case, the pragmatics may be able to further refine/disambiguate the correct meaning in the specific context.

To determine the direction/dependency, we use the notion of *sense extension* (Copestake and Briscoe, 1995), in which the secondary sense is produced from the primary sense to yield a consistent interpretation of the word in a specific context. As an example, for the noun **chicken**, ANIMAL meaning is considered as the primary and MEAT as the secondary, because meat is obtained from an eatable animal (i.e., *animal-grinding* lexical rule (Copestake and Briscoe, 1992)). Similarly for the verb **write**, CREATION is considered as the primary and COMMUNICATION as the secondary, because communication takes place through the (implicit) object that is just created.

In addition to sense extension, we use another criteria, namely, the ontological generality of the basic types. This is because a transitive inference through a ontologically general type may result in a relation where unrelated (specific) types are combined. For example, given three types τ, σ, η where σ is more general than τ and η, suppose there are two systematic relations R and S such that $\tau R \sigma$ and $\sigma S \eta$. Then we have compositions $(\tau R \sigma) S \eta$ and $\tau R(\sigma S \eta)$ through transitive inference. But here, since σ is more general than τ and η (i.e., the greatest lower bound), its existence in the compositions adds no information (thus $\tau R \sigma \equiv \tau$ and $\sigma S \eta \equiv \eta$). Thus those compositions are essentially reduced to a relation between τ and η ($\tau R \eta$ or $\tau S \eta$), which may not be systematic in some cases, particularly when τ and η are different domain categories. In order to prevent this kind of compositions, we restricted the direction of a systematic relation to be from the specific type to the general type, if one of the member types is the generalization of the other.

3 Induction Method

Our induction method is semi-automatic, requiring a manual filtering step between the phased automatic processing. The basic scheme of our method is to first identify the prominent pair-wise cooccurrence between any two basic types (abstract senses), and then build more complex types (underspecified classes) by the composition of those cooccurrences. But instead of generating/composing all possible types statically, we only maintain the pair-wise relations in a graph representation called *type dependency graph*, and dynamically form/induce the underspecified classes during the phase when each WordNet entry is assigned the class label(s).

Based on the definitions and assumptions described in the previous section 2, underspecified semantic classes are induced from WordNet 1.6 (released December 1997) by the following steps:

1. Select a set of abstract (coarse-grained) senses from WordNet taxonomies as basic semantic types. This step is done manually, to determine the right level of abstraction to capture systematic polysemy.

2. Create a *type dependency graph* from ambiguous words in WordNet. This step is done by two phased analyses: an automatic analysis followed by a manual filtering.

3. Generate a set of underspecified semantic classes by partitioning the senses of each word into a set of basic types. Each set becomes an underspecified semantic class. This step is fully automatic.

Each step is described in detail below.

3.1 Coarse-grained Basic Types

As has been pointed out in many previous literature, there are many regularities between polysemous senses, and these regularities seem to hold across words. For example, words such as **chicken** and **duck** which have ANIMAL sense often have MEAT meaning also. This regularity holds at an abstract level rather than the word sense level. Therefore, the first step in the induction is to select a set of abstract senses that are useful in capturing the regularity. To this end, WordNet is a good resource because word senses (or *synsets*) are organized in taxonomies.

Ideally, basic types should be semantically orthogonal, to function essentially as the "axes" in a high-dimensional semantic space. Good candidates would be the top abstract nodes in the WordNet taxonomies or lexicographers' file names listed in the sense entries. However, both of them fall short of forming a set of orthogonal axes because of several reasons. First, some categories are ontologically more general than others (eg. `change` category in verbs). As we discussed in the previous section 3, this property is problematic. Second, domain categories are mixed in with ontological categories (eg. `competition` or `body` versus `motion` categories in the verbs). Third, particularly for the verbs, senses that seem to take different argument noun types are found under the same category (eg. "ingest" and "use" in `consumption` category). Therefore, some WordNet categories are broken into more specific types.

For the verbs, the following 18 abstract basic types are selected.

change(CHA)	communication(COMM)	cognition(COG)
contact(CONT)	motion(MOT)	competition(COMP)
emotion(EMO)	perception(PER)	possession(POSS)
stative(STA)	social(SOC)	weather(WEA)
ingestion(ING)	use(USE)	body(BOD)
phy_creation(PCR)	mental_creation(MCR)	verbal_creation (VCR)

These are mostly taken from the classifications made by lexicographers. Two classes ("consumption" and "creation" are subdivided into finer categories (`ingestion`, `use` and `physical/mental/verbal_creation`, respectively) according to the different predicate-argument structures they take.

For the nouns, 31 basic types are selected from WordNet top categories (unique beginners):[4]

entity(ENT)	life_form(LIF)	causal_agent(AGT)
human(HUM)	animal(ANI)	plant(PLA)
object(OBJ)	natural_object(NOBJ)	substance(SUB)
food(FOOD)	artifact(AFT)	article(ART)
location(LOC)	psych_feature(PSY)	cognition(COG)
feeling(FEEL)	motivation(MOT)	abstraction(ABS)
time(TIME)	space(SPA)	attribute(ATT)
relation(REL)	social_relation(SREL)	communication(COMM)
shape(SHA)	measure(MEA)	event(EVE)
action(ACT)	possession(POSS)	state(STA)
phenomena(PHE)		

Senses under the lexicographers' class "group" are redirected to other classes, assuming a collection of a type has the same basic semantic properties as the individual type.

[4]Noun top categories in WordNet do not match exactly with lexicographers' file names. In our present work, noun types are determined by actually traversing the hierarchies, therefore they correspond to the top categories.

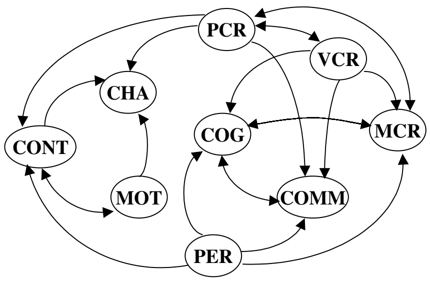

Figure 1: Part of type dependency graph for WordNet verbs

3.2 Type Dependency Graph

After the basic types are selected, the next step is to create a *type dependency graph*: a directed graph in which nodes represent the basic types, and directed edges correspond to the systematic relations between two basic types.

The type dependency graph is constructed by an automatic statistical analysis followed by a manual filtering process, as described below. The premise behind the automatic analysis here is that, if there is a systematic relation between two types, and if the regularity is prominent, it can be captured by the type cooccurrence statistics. In machine learning, several statistical techniques have been developed which discover dependencies among features (or causal structures), such as Bayesian network learning (Spirtes *et al.*, 1993). Those techniques use sophisticated methods that take into consideration of multiple antecedents/causations and so on, and build a complex and precise model with probabilities associated with edges. In our present work however, WordNet is compiled from human lexicographers' entries, thus the data has a fair amount of arbitrariness (i.e., noisy data). Therefore, we chose a simple technique which yields a simpler network, and used the result as a rough approximation of the type dependencies. Then, this result is manually corrected by filtering out the spurious relations (i.e., false positives) and augmenting with the missing ones (i.e., false negatives). This scheme will effectively yield results that reflect our prior knowledge about WordNet which the statistical analysis did not assume. It also helps discover the relations we have not realized with respect to the basic types selected from the previous step.

The type dependency graph is constructed in the following way. First, for all type-pairs extracted from the ambiguous words in WordNet, *mutual information* is computed to obtain the association by using the standard formula: for type $t1$, $t2$, a mutual information $I(t1,t2)$ is

$$I(t1,t2) = lg \ \frac{\frac{f(t1 \wedge t2)}{N}}{\frac{f(t1)}{N} \times \frac{f(t1)}{N}}$$

where $f(t)$ is the number of occurrence of the type t, and N is the size of the data. The association between two types are considered prominent when the mutual information value was greater than some threshold (in our current implementation, it is 0).

At this point, type associations are still undirected because mutual information is commutative (i.e., symmetric). Then, these associations are manually inspected to create a directed type dependency graph in the next phase. The manual filtering does two things: to prune the false positives and add back the false negatives, and to determine the direction of the correct associations. Detected false positives are mostly homonyms (including metaphors) (eg. `WEA-EMO` (`weather` and `emotion`) verb type pair for words such as the word **ignite**). False negatives are mostly the ones that we know exist, but were not significant according to the cooccurrence statistics (eg. `ANI-FOOD` in nouns). As a heuristic to detect the false negatives, we used the cross-categorical inheritance in the taxonomies in which category switches as the hierarchy is traversed up. The overall accuracy of the automatic analysis was, for the verbs, out of 156 possible type pairs (from 18 basic types; $\frac{18 \times 17}{2} = 156$ pairs), 116 pairs were determined prominent by mutual information. There were 18 false positives and 10 false negatives, yielding 91 % recall and 84 % precision. For the nouns, out of 465 possible type pairs (from 31 basic types; $\frac{31 \times 30}{2} = 465$ pairs), 111 pairs were determined prominent by mutual information. There were 27 false positives and 5 false negatives, yielding 94 % recall and 76 % precision. For both categories, recall seems relatively high.

The direction of the associations are determined by the way described in section 2. Note for some associations which involve equally general/specific types ontologically (such as `COG (cognition)` and `COMM (communication)`), the direction was considered bidirectional (unless sense extension strongly suggests the dependency).

A part of the type dependency graph for WordNet verbs is shown in Figure 1. As you can see, directions of edges are determined by the assumptions described in the previous section 2.2. For instance, `CHA (change)` has no outgoing edges since it is ontologically general. On the other hand, `PER (perception)` has no incoming edges since it is domain-specific. Sense extension is observed from `PCR` to `COMM`, which would be given an interpretation "an creation act (which resulted some object) performed for communication purpose (through the object)".

3.3 Underspecified Semantic Classes

Underspecified semantic classes are automatically formed by partitioning the ambiguous senses of each word according to the type dependency graph.

Using the type dependency graph, all words in WordNet verb and noun categories are assigned one or more type *partitions*. A partition is an ordered set of basic types (abstracted from the fine-grained word senses in the first step) keyed by the primary type emcompassing the secondary types. From a list of frequency-ordered senses of a WordNet word, a partition is created by taking one of the three most frequent types (listed as the first three senses in the WordNet entry) as the primary and collecting the secondary types from the remaining list according to the type dependency graph.[5] Here, the secondary types are taken only from the nodes/types that are directly connected to the primary type. That is because we assumed if an indirect transitive dependency of $t1$ on $t3$ through $t2$ is strong enough, it will be captured as a direct dependency. This scheme also ensures the existence of a core concept in every partition (thus more implausible than transitive composition). Then, if a partition can be made that covers all types/senses, the word is considered systematically polysemous, and the partition is taken as an underspecified semantic class.

As an example, for the verb **write** whose sense list is `(VCR COMM PCR CHA)`,[6] the first 3 types `VCR`, `COMM` and `PCR` are picked in turn as the primary type to see whether a partition can be created that encompasses all other member types. In this case, a partition keyed by `PCR` can cover all member types (see the type dependency graph in Figure 1), thus a class `VCR-COMM-PCR-CHA` is created. The systematic relation of this class would be "a change or creation action which involves words (and resulted some object), performed for communication purpose (through the

[5] The reason we look at the first three senses is because primary types are not always listed as the most frequent sense in the WordNet sense lists (or in actual usage for that matter). We chose the first three senses because the average degree of polysemy is around 3 for WordNet (version 1.6) verbs and nouns.

[6] The original 9 senses listed in WordNet were compressed down to these 4 basic types.

Table 1: Example verbs in CONT classes

Verb Class	*Verbs*
CONT-CHA	*blend, crush, enclose, fasten, fold, puncture, tie, weld*
CONT-MOT	*beat, chop, fumble, jerk, kick, press, spread, whip*
CONT-POSS	*pluck, release, seize, sponge*
CONT-MOT-CHA	*dip, gather, mount, take_out*
CONT-MOT-POSS	*carry, cover, fling, toss*

Table 2: Underspecified classes and systematic polysemies in WordNet

Category	Basic Types	Underspecified Classes	Poly Words Covered	Poly Words in WordNet	(%)
verb	18	136	1,476	4,567	32
noun	31	325	4,012	12,564	32

object)".

Here, we must note here that this partitioning technique can be used to divide the senses of homonyms. A homonym is a word which has more than one unrelated senses/types. So, if the senses of a word do not conform to a single partition (thus the word is a homonym), the procedure is applied recursively to find all partitions that cover the sense list exhaustively. In the experiment, we actually process this way, and assigned (one or more) classes to all WordNet verbs and nouns.[7]

For the verbs and nouns in WordNet 1.6, 136 underspecified verb classes and 325 underspecified noun classes are formed. Some verbs of the classes involving `contact (CONT)` are shown in Table 1. We can observe from the words assigned to each class that member types are indeed systematically related. For example, CONT-MOT class represents an action which involves physical contact resulting from motion (MOT). Words assigned to this class do seem to have `motion` flavor. On the other hand, CONT-POSS class represents a transfer of possession (POSS) which involves physical contact. Again, words in this class do seem to be used in a context in which possession of something is changed. For the more polysemous class CONT-MOT-POSS, words in this class, for instance **toss**, do seem to cover all three member types.

4 Discussion

To assess the results, our 136 verb classes cover 1,476 polysemous verbs, and 325 noun classes cover 4,012 polysemous nouns in WordNet. Table 2 shows the summary of our results. For both verbs and nouns, our underspecified classes cover 32 % of the polysemous words in WordNet 1.6.

These results, however, can not be compared directly with CORELEX, which consists of

[7] As a note, by using the underspecified classes, the average number of senses for a word in WordNet decreased from 3.57 to 2.39 for the verbs, and from 2.73 to 2.21 for the nouns (both excluding unambiguous words).

126 underspecified noun classes (compressed from 529 or 325 classes) composed of 39 basic types, and covers 39,937 nouns out of 42,116 noun stems. That is because there are several basic differences in the implementation of our method and CORELEX, including the version of WordNet used (1.6 in ours versus 1.5 in CORELEX), and the choice of basic types (31 in ours versus 39 in CORELEX). In particular, the taxonomies in the two versions of WordNet are slightly but sufficiently different enough to affect the assignment of basic types to each WordNet entry. Also, the set of basic types in our current implementation do not map exactly to those in CORELEX (although there seem to be a large overlap between the two sets). However, by looking at the number of words assigned to each class and inspecting the actual words classified in CORELEX, we speculate a large portion of the discrepancy of the coverage (by which we mean our classes cover a smaller set of nouns) seems to come from the difference of the way CORELEX basic types are assigned to the words with multiple inheritance for some familiar classes (such as `human`, which is inherited from `agent` and `life_form`), and the words which are assigned a singleton class (such as the class `'act'` which consists of one basic type `act`). In our method, we counted those words as a monosemy since those words have only one sense listed in the WordNet entry. Thus they are not in the original 12,564 polysemous nouns accounted in WordNet 1.6, which we applied our induction method to. But since there are many such words in CORELEX, those words seem make up a large portion of the discrepancy of coverage.

However, we also realize that our induced classes may be incomplete. In other words, our method did not find/create all the underspecified classes that could have been induced. This is due to some fundamental difficulties of our method. Our method creates underspecified classes by composing the binary relations that hold between two basic types. Therefore, every two types in a composition are in a systematic relation by our method. This bottom-up construction scheme is rigorous, but it never creates an n-ary composite relation (where $n > 2$) in which some of the base binary relations are not prominent when they are taken individually (thus no edge exists for those relations in the type dependency graph). However, such an n-ary relation may be systematic, mainly when the non-prominent basic relations are indeed systematic, but were not determined as such in the type dependency graph because of the seeming low *productivity* (in the same sense as in lexical rules) or applicability in the general domain. Thus, those n-ary systematic relations were missed in our method, but could be selected in the manually constructed CORELEX.

Another difficulty in our method which may cause the incompleteness is that transitive inference is not allowed (i.e., composition is done only through one single core type, as we described earlier in section 3.3). The objective of this restriction is to decrease the possilibity of producing implausible compositions. In as much as it works for that purpose, it also restricts the kinds of systematic relations the induction method can find. Consequently our method is rather conservative in inducing underspecified classes.

As for the manual effort required, CORELEX was created by examining each of the polysemous type combinations of length greater than 1 and to which more than one word was assigned (529 classes were reported in (Buitelaar, 1998)). Each examination (i.e., determination of systematicity) involves deriving the decomposition of the member types and hypothesizing the context in which the decomposition yields some interpretation. On the other hand, our method requires inspecting all pair-wise combinations of basic types (465 noun type pairs in our current implementation). The difference of the actual number of combinations that need examination depends on the content of the resource (i.e., coverage and degree of ambiguity) for CORELEX, and the number of basic types for our method. However, each examination in our method is much simpler because the decomposition is only binary: either $\tau_1 \circ \tau_2$ or $\tau_1 \bullet \tau_2$.

More importantly, our method utilizes the result of the automatic analysis (mutual information) in the manual examination, therefore the determination of systematicity is more informed. In other words, by knowing the strength of the association between two types that was extracted from the actual content of the (large) data, and with respect to the manually selected basic types, the determination can become more assured and consistent than by judging with mere intuition. This is another important advantage from applying the automatic analysis in our method.

Lastly, the induction method described in this paper should be considered as an initial attempt to automatically acquiring systematic polysemy from a broad-coverage lexical resource. The task is essentially to map our semantic/ontological knowledge about the systematicity of word meanings to some computational terms for a given lexical resource. In our present work, we mapped the systematicity to the cooccurrence of word senses. But the mapping only by computational/automatic means (mutual information) was not possible: manual filtering was further needed to enhance the mapping.

5 Future Work

The work described in this paper is still preliminary. We must further analyze the current results and assess the the accuracy/applicability of the induced classes (i.e, systematic relation may not hold for all words in the class; it may only be a spurious cooccurrence of senses).

Then our next plan is to decrease the degree of manual intervention and further automate the induction process. The first two steps of the induction, to select basic types and to create type dependency graph, could be improved, possibly to a fully automated process. The issues, then, will include how to derive the right level of abstraction for broad domains (eg. Hearst and Schütze, 1993; Cucchiarelli and Velardi, 1997), and how to detect and distinguish systematic polysemies from homonyms, and how to incorporate our linguistic knowledge about the word sense systematicity as a prior domain knowledge in the induction algorithm for the given resource (WordNet).

Another important extension is to define a representation for each underspecified class that explicitly encodes how the senses relate to one another, possibly through complex interactions between different aspects of the object/action which may involve implicit arguments. Then, it will be interesting to see such representation defined at the abstract polysemous class level can be combined with micro (word sense) level representation (eg. (Harabagiu and Moldovan, 1997)).

Acknowledgments

The author would like to thank Paul Buitelaar for helpful discussions, insights and encouragement.

References

Buitelaar, P. (1997). A Lexicon for Underspecified Semantic Tagging, In *Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*. pp. 25-33.

Buitelaar, P. (1998). CORELEX: Systematic Polysemy and Underspecification. Ph.D. dissertation, Department of Computer Science, Brandeis University.

Copestake, A. and Briscoe, T. (1992). Lexical Operations in a Unification-based Framework, In *Lexical Semantics and Knowledge Representation*, J. Pustejovsky (eds.), pp. 101-119, Springer-Verlag.

Copestake, A. and Briscoe, T. (1995). Semi-productive Polysemy and Sense Extension. *Journal of Semantics*, **12**.

Cucchiarelli, A. and Velardi, P. (1997). Automatic Selection of Class Labels from a Thesaurus for an Effective Semantic Tagging of Corpora, In *Proceedings of the 5th conference on Applied Natural Language Processing* (ANLP-97), pp. 380-387.

Harabagiu, S. and Moldovan, D. (1997). TextNet – A Text-based Intelligent System. *Natural Language Engineering*, **3**.

Hearst, M. and Schütze, H. (1993). Customizing a Lexicon to Better Suit a Computational Task, In *Proceedings of the ACL SIGLEX Workshop on Lexical Acquisition from Text*, Columbus, OH.

Kilgarriff, A. (1992). Polysemy. Ph.D. Dissertation, University of Sussex.

Kilgarriff, A. (1997). Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction, In *Proceedings of the International Workshop on Lexically Driven Information Extraction.*

Miller, G. (eds.) (1990). WORDNET: An Online Lexical Database. *International Journal of Lexicography*, **3** (4).

Ng, H.T., and Zelle, J. (1997). Corpus-based Approaches to Semantic Interpretation in Natural Language Processing. *AI Magazine*, 18(4).

Ostler, N. and Atkins, B. (1992). Predictable Meaning Shift: Some Linguistic Properties of Lexical Implication Rules, In *Lexical Semantics and Knowledge Representation*, J. Pustejovsky (eds.), pp. 87-100, Springer-Verlag.

Pustejovsky, J. (1995). *The Generative Lexicon*, The MIT Press.

Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction and Search*, Springer-Verlag.

Verspoor, C. (1997). Contextually-dependent Lexical Semantics. Ph.D. dissertation, University of Edinburgh.