Automated Induction of Sense in Context

James PUSTEJOVSKY, Patrick HANKS, Anna RUMSHISKY
Brandeis University
Waltham, MA 02454, USA
{jamesp,patrick,arum}@cs.brandeis.edu

1 Introduction

In this work, we introduce a model for sense assignment which relies on assigning senses to the contexts within which words appear, rather than to the words themselves. We argue that word senses as such are not directly encoded in the lexicon of the language. Rather, each word is associated with one or more stereotypical syntagmatic patterns, which we call *selection contexts*. Each selection context is associated with a meaning, which can be expressed in any of various formal or computational manifestations. We present a formalism for encoding contexts that help to determine the semantic contribution of a word in an utterance. Further, we develop a methodology through which such stereotypical contexts for words and phrases can be identified from very large corpora, and subsequently structured in a selection context dictionary, encoding both stereotypical syntactic and semantic information. We present some preliminary results.

2 CPA Methodology

The Corpus Pattern Analysis (CPA) technique uses a semi-automatic bootstrapping process to produce a dictionary of selection contexts for predicates in a language. Word senses for verbs are distinguished through corpus-derived syntagmatic patterns mapped to Generative Lexicon Theory (Pustejovsky (1995)) as a linguistic model of interpretation, which guides and constrains the induction of senses from word distributional information. Each pattern is specified in terms of lexical sets for each argument, shallow semantic typing of these sets, and other syntagmatically relevant criteria (e.g., adverbials of manner, phrasal particles, genitives, negatives).

The procedure consists of three subtasks: (1) the manual discovery of selection context patterns for specific verbs; (2) the automatic recognition of instances of the identified patterns; and (3) automatic acquisition of patterns for unanalyzed cases. Initially, a number of patterns are manually formulated by a lexicographer through corpus pattern analysis of about 500 occurrences of each verb lemma. Next, for higher frequency verbs, the remaining corpus occurrences are scrutinized to see if any low-frequency patterns have been missed. The patterns are then translated into a feature matrix used for identifying the sense of unseen instances for a particular verb. In the remainder of this section, we describe these subtasks in more detail. The following sections explain the current status of the implementation of these tasks.

2.1 Lexical Discovery

Norms of usage are captured in what we call selection context patterns. For each lemma, contexts of usage are sorted into groups, and a stereotypical CPA pattern that captures the relevant semantic and syntactic features of the group is recorded. For example, here is the set of common patterns for the verb *treat*.

(1) CPA pattern set for `treat`:

```
I. [[Person 1]] treat [[Person 2]] ({at | in} [[Location]])
   (for [[Event = Injury | Ailment]]); NO [Adv[Manner]]
II. [[Person 1]] treat [[Person 2]] [Adv[Manner]]
IIIa. [[Person]] treat [[TopType 1]] {{as | like} [[TopType 2]]}
IIIb. [[Person]] treat [[TopType]] {{as if | as though | like}
    [CLAUSE]}
IV. [[Person 1]] treat [[Person 2]] {to [[Event]]}
V. [[Person]] treat [[PhysObj | Stuff 1]] (with [[Stuff 2]])
```

There may be several patterns realizing a single sense of a verb, as in (IIIa/IIIb) above. Additionally, many patterns have alternations, recorded in satellite CPA patterns. Alternations are linked to the main CPA pattern through the same sense-modifying mechanisms as those that allow for coercions to be understood. However, alternations are different realizations of the same norm. For example, the following are alternations for `treat`, pattern (I):

```
[[Person 1 <--> Medicament | Med-Procedure | Institution]]
[[Person 2 <--> Injury | Ailment | Bodypart]]
```

CPA Patterns

A CPA pattern extends the traditional notion of selectional context to include a number of other contextual features, such as minor category parsing and subphrasal cues. Accurate identification of the semantically relevant aspects of a pattern is not an obvious and straightforward procedure, as has sometimes been assumed in the literature. For example, the presence or absence of an adverbial of manner in the third valency slot around a verb can dramatically alter the verb's meaning. Simple syntactic encoding of argument structure, for instance, is insufficient to discriminate between the two major senses of the verb *treat*, as illustrated below.

(3) a. They say their bosses **treat** them with respect.
b. Such patients are **treated** with antibiotics.

The ability to recognize the shallow semantic type of a phrase in the context of a predicate is of course crucial —for example, in (3a) recognizing the PP as (a) an adverbial, and (b) an adverbial of manner, rather than an instrumental co-agent (as in (3b)), is crucial for assigning the correct sense to the verb *treat* above.

There are four constraint sets that contribute to the patterns for encoding selection contexts. These are:

(4) a. *Shallow Syntactic Parsing*: Phrase-level recognition of major categories.
b. *Shallow Semantic Typing*: 50-100 primitive shallow types, such as Person, Institution, Event, Abstract, Artifact, Location, and so forth. These are the top types selected from the Brandeis Shallow Ontology (BSO), and are similar to entities (and some relations) employed in Named Entity Recognition tasks, such as TREC and ACE.
c. *Minor Syntactic Category Parsing*: e.g., locatives, purpose clauses, rationale clauses, temporal adjuncts.
d. *Subphrasal Syntactic Cue Recognition*: e.g., genitives, partitives, bare plural/determiner distinctions, infinitivals, negatives.

The notion of a selection context pattern, as produced by a human annotator, is expressed as a BNF specification in Table 1.[1] This specification relies on word order to specify argument position, and is easily translated to a template with slots allocated for each argument. Within this grammar, a semantic roles can be specified for each argument, but this information currently is not used in the automated processing.

Brandeis Shallow Ontology

The Brandeis Shallow Ontology (BSO) is a shallow hierarchy of types selected for their prevalence in manually identified selection context patterns. At the time of writing, there are just 65 types, in terms of which patterns for the first one hundred verbs have been analyzed. New types are added occasionally, but only when all possibilities of using existing types prove inadequate. Once the set of manually extracted patterns is sufficient, the type system will be re-populated and become pattern-driven.

The BSO type system allows multiple inheritance (e.g. `Document` $\sqsubseteq$ `PhysObj` and `Document` $\sqsubseteq$ `Information`. The types currently comprising the ontology are listed below. The BSO contains type assignments for 20,000 noun entries and 10,000 nominal collocation entries.

[1] Round brackets indicate optional elements of the pattern, and curly brackets indicate syntactic constituents.

TOPTYPE	Location
Event	Dwelling
Action	Accommodation
SpeechAct	*Energy*
Activity	*Abstract*
Process	Attitude
State	Emotion
Entity	Responsibility
PhysObj	Privilege
Artifact	Rule
Machine	Information
Vehicle	Document
Hardware	Music
Document	Artwork
Music	Film
Artwork	Program
Film	Software
Program	Word
Software	Language
Medium	Concept
Garment	Property
Drug	VisibleFeature
Substance	Color
Vapor	Shape
Animate	TimePeriod
Bird	Holiday
Horse	CourseOfStudy
Person	Cost
Human Group	Asset
Plant	Route
PlantPart	
Body	
BodyPart	
Institution	
HumanGroup	

Corpus-driven Type System

The acquisition strategy for selectional preferences for predicates proceeds as follows:

(5) a. Partition the corpus occurrences of a predicate according to the selection contexts pattern grammar, distinguished by the four levels of constraints mentioned in (4). These are uninterpreted patterns for the predicate.

b. Within a given pattern, promote the statistically significant literal types from the corpus for each argument to the predicate. This induces an interpretation of the pattern, treating the promoted literal type as the specific binding of a shallow type from step (a) above.

c. Within a given pattern, coerce all lexical heads in the same shallow type for an argument, into the promoted literal type, assigned in (b) above. This is a coercion of a lexical head to the interpretation of the promoted literal type induced from step (b) above.

In a sense, (5a) can be seen as a broad multi-level partitioning of the selectional behavior for a predicate according to a richer set of syntactic and semantic discriminants. Step (5b) can be seen as capturing the norms of usage in the corpus, while step (5c) is a way of modeling the exploitation of these norms in the language (through coercion, metonymy, and other generative operations). To illustrate the way in which CPA discriminates uninterpreted patterns from the corpus, we return to the verb *treat* as it is used in the BNC. Two of its major senses, as listed in (1), emerge as correlated with two distinct context patterns, using the discriminant constraints mentioned in (4) above.

CPA-Pattern → Segment verb-lit Segment | verb-lit Segment | Segment verb-lit | CPA-Pattern ’;’ Element
Segment → Element | Segment Segment | ’’ Segment ’’ | ’(’ Segment ’)’ | Segment ’|’ Segment
Element → literal | ’[’ Rstr ArgType ’]’ | ’[’ Rstr literal ’]’ | ’[’ Rstr ’]’ | ’[’ NO Cue ’]’ | ’[’ Cue ’]’
Rstr → POS | Phrasal | Rstr ’|’ Rstr | epsilon
Cue → POS | Phrasal | AdvCue
AdvCue → ADV ’[’ AdvType ’]’
AdvType → Manner | Dir | Location
Phrasal → OBJ | CLAUSE | VP | QUOTE
POS → ADJ | ADV | DET | POSDET | COREF POSDET | REFL-PRON | NEG |
MASS | PLURAL | V | INF | PREP | V-ING | CARD | QUANT | CONJ
ArgType → ’[’ SType ’]’ | ’[’ SType ’=’ SubtypeSpec ’]’ | ArgType ’|’ ArgType | ’[’ SType ArgIdx ’]’ |
’[’ SType ArgIdx ’=’ SubtypeSpec ’]’
SType → AdvType | TopType | Entity | Abstract | PhysObj | Institution | Asset | Location | Human | Animate |
Human Group | Substance | Unit of Measurement | Quality | Event | State of Affairs | Process
SubtypeSpec → SubtypeSpec ’|’ SubtypeSpec | SubtypeSpec ’&’ SubtypeSpec | Role | Polarity | LSet
Role → Role | Role ’|’ Role | Benficiary | Meronym | Agent | Payer
Polarity → Negative | Positive
LSet → Worker | Pilot | Musician | Competitor | Hospital | Injury | Ailment | Medicament | Medical Procedure |
Hour-Measure | Bargain | Clothing | BodyPart | Text | Sewage | Part | Computer | Animal
ArgIdx → <number>
literal → word
CARD → <number>
POSDET → my | your | ...
QUANT → CARD | a lot | longer | more | many | ...
verb-lit → <verb-word-form>
word → <word>
NEG → not
INF → to

Table 1: Pattern grammar

(6) a. `[[Person 1]] treat [[Person 2]]; NO [Adv[Manner]]`
b. `[[Person 1]] treat [[Person 2]] [Adv[Manner]]`

Given a distinct (contextual) basis on which to analyze the actual statistical distribution of the words in each argument position, we can promote statistically relevant and significant literal types for these positions. For example, for pattern (a) above, this induces `Doctor` as Person 1, and `Patient` as bound to Person 2. This produces the interpreted context pattern for this sense as shown below.

(7) `[[doctor]] treat [[patient]]`

Promoted literal types are corpus-derived and predicate-dependent, and are syntactic heads of phrases that occur with the greatest frequency in argument positions for a given sense pattern; they are subsequently assumed to be subtypes of the particular shallow type in the pattern. Step (5c) above then enables us to bind the other lexical heads in these positions as *coerced* forms of the promoted literal type. This can be seen below in the concordance sample, where *therapies* is interpreted as `Doctor`, and *people* and *girl* are interpreted as `Patient`.

(8) a. *a doctor* who **treated** *the girl* till an ambulance arrived.
b. *over 90,000 people* have been **treated** for *cholera*
c. *nonsurgical therapies* to **treat** *the breast cancer*, which

Model Bias

The assumption within GL is that semantic types in the grammar map systematically to default syntactic templates (cf. Pustejovsky (1995)). These are termed *canonical syntactic forms (CSFs)*. For example, the CSF for the type `proposition` is a tensed S. There are, however, many possible realizations (such as infinitival S and NP) for this type due to the different possibilities available from generative devices in a grammar, such as coercion and co-composition. The resulting set of syntactic forms associated with a particular semantic type is called a *phrasal paradigm* for that type. The model bias provided by GL acts to guide the interpretation of purely statistically based measures.

2.2 Automatic Recognition of Pattern Use

Essentially, this subtask is similar to the traditional supervised WSD problem. Its purpose is (1) to test the discriminatory power of CPA-derived feature-set, (2) to extend and refine the inventory of features captured by the CPA patterns, and (3) to allow for predicate-based argument groupings by classifying unseen instances. Extension and refinement of the inventory of features should involve feature induction, but at the moment this part has not been implemented. During the lexical discovery stage, lexical sets that fill some of the argument slots in the patterns are instantiated from the training examples. As more predicate-based lexical sets within shallow types are explored, the data will permit identification of the types of features that unite elements in lexical sets.

2.3 Automatic Pattern Acquisition

The algorithm for automatic pattern acquisition involves the following steps:

(9) a. Collect all constituents in a particular argument position;
b. Identify syntactic alternations;
c. Perform clustering on all nouns that occur in a particular argument position of a given predicate;
d. For each cluster, measure its relatedness to the known lexical sets, obtained previously during the lexical discovery stage and extended through WSD of unseen instances. If none of the existing lexical sets pass the distance threshold, establish the cluster as a new lexical set, to be used in future pattern specification.

Step (9d) must include extensive filtering procedures to check for shared semantic features, looking for commonality between the members. That is, there must be some threshold overlap between subgroups of the candidate lexical set and and the existing semantic classes. For instance, checking if, for a certain percentage of pairs in the candidate set, there already exists a set of which both elements are members.

3 Current Implementation

The CPA patterns are developed using the British National Corpus (BNC). The sorted instances are used as a training set for the supervised disambiguation. For the disambiguation task, each pattern is translated into into a set of preprocessing-specific features.

The BNC is preprocessed with the RASP parser and semantically tagged with BSO types. The RASP system (Briscoe and Carroll (2002)) generates full parse trees for each sentence, assigning a probability to each parse. It also produces a set of grammatical relations for each parse, specifying the relation type, the headword, and the dependent element. All our computations are performed over the single top-ranked tree for the sentences where a full parse was successfully obtained. Some of the RASP grammatical relations are shown in (10).

(10) **subjects**: ncsubj, clausal (csubj, xsubj)
objects: dobj, iobj, clausal complement
modifiers: adverbs, modifiers of event nominals

We use endocentric semantic typing, i.e., the headword of each constituent is used to establish its semantic type. The semantic tagging strategy is similar to the one described in Pustejovsky et al. (2002), and currently uses a subset of 24 BSO types.

A CPA pattern is translated into a feature set, currently using binary features. It is further complemented with other discriminant context features which, rather than distinguishing a particular pattern, are merely *likely* to occur with a given subset of patterns; that is, the features that only partially determine or co-determine a sense. In the future, these should be learned from the training set through feature induction from the training sample, but at the moment, they are added manually. The resulting feature matrix for each pattern contains features such as those in (11) below. Each pattern is translated into a template of 15-25 features.

(11) **Selected context features:**
a. `obj_institution`: object belongs to the BSO type 'Institution'
b. `subj_human_group`: subject belongs to the BSO type 'HumanGroup'
c. `mod_adv_ly`: target verb has an adverbial modifier, with a -ly adverb
d. `clausal_like`: target verb has a clausal argument introduced by 'like'
e. `iobj_with`: target verb has an indirect object introduced by 'with'
f. `obj_PRP`: direct object is a personal pronoun
g. `stem_VVG`: the target verb stem is an -ing form

Each feature may be realized by a number of RASP relations. For instance, a feature dealing with objects would take into account RASP relations `'dobj'`, `'obj2'`, and `'ncsubj'` (for passives).

4 Results and Discussion

The experimental trials performed to date are too preliminary to validate the methodology outlined above in general terms for the WSD task. Our results are encouraging however, and comparable to the best performing systems reported from Senseval 2. For our experiments, we implemented two machine learning algorithms, instance-based k-Nearest Neighbor, and a decision tree algorithm (a version of ID3). Table 2 shows the results on a subset of verbs that have been processed, also listing the number of patterns in the pattern set for each of the verbs.[2]

verb	number of patterns	training set	accuracy	
			ID3	kNN
edit	2	100	87%	86%
treat	4	200	45%	52%
submit	4	100	59%	64%

Table 2: Accuracy of pattern identification

Further experimentation is obviously needed to adequately gauge the effectiveness of the selection context approach for WSD and other NLP tasks. It is already clear, however, that the traditional sense enumeration approach, where senses are associated with individual lexical items, must give way to a model where senses are assigned to the contexts within which words appear. Furthermore, because the variability of the stereotypical syntagmatic patterns that are associated with words appears to be relatively small, such information can be encoded as lexically-indexed contexts. A comprehensive dictionary of such contexts could prove to be a powerful tool for a variety of NLP tasks.

References

T. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002).*

J. Pustejovsky, A. Rumshisky, and J. Castano. 2002. Rerendering Semantic Ontologies: Automatic Extensions to UMLS through Corpus Analytics. In *LREC 2002 Workshop on Ontologies and Lexical Knowledge Bases.*.

J. Pustejovsky. 1995. *Generative Lexicon.* Cambridge (Mass.): MIT Press.

[2] Test set size for each lemma is 100 instances, selected out of several randomly chosen segments of BNC, non-overlapping with the training set