Data Sparseness and the Acquisition of Syntax-Semantics Mappings from Corpora

Christopher Hogan and Lori S. Levin
Carnegie Mellon University
chogan@nl.cs.cmu.edu, lsl@cs.cmu.edu

Abstract: This paper discusses the acquisition of syntax-to-semantics mappings from corpora, concentrating on two major obstacles: data sparseness and verb class flexibility. We report on a preliminary experiment using the Susanne corpus that confirms the severity of the sparseness problem. We then outline a program for acquiring syntax-to-semantics mappings using a corpus and a pre-existing ontology.

1 Introduction

Currently, many computational linguists have undertaken to assemble large machine-readable dictionaries for the purpose of parsing and understanding unrestricted text. Although there are techniques for converting preexisting on-line dictionaries for this purpose, conventional dictionaries are inadequate because, with some exceptions such as COBUILD, they are biased toward uncommon usage and away from common usage (Church and Mercer, [CM93]). Interest has therefore turned to the large text corpora that now exist on-line. This paper will discuss the acquisition of syntax-to-semantics mappings from corpora. We consider syntax-to-semantics mappings to be associations of syntactic arguments with semantic roles as shown in parentheses below each sentence in (1).

(1) a. The teacher gave books to the students.
(agent=subject, theme=object, goal=oblique)

b. The teacher gave the students books.
(agent=subject, goal=object, theme=2nd object)

c. Books were given to the students by the teacher.
(theme=subject, goal=oblique, agent=oblique)

d. The students were given books by the teacher.
(goal=subject, theme=2nd object, agent=oblique)

In dealing with syntax-to-semantics mappings our work is broader in scope than much recent work on extracting only syntactic subcategorization frames from corpora (e.g., Brent ([Bre93]), Manning ([Man93]), and others). In addition, we will focus on the issue of data sparseness. We show preliminary results that suggest the near impossibility of ever acquiring complete information about subcategorization. We argue that robust NLP for unrestricted text will have to be based on partial information about subcategorization with probabilistic procedures for assigning likely meanings to previously unseen subcategorization frames. Although our work on dealing with data sparseness is at a very preliminary stage, we will present the direction a solution might take.

Current corpus-based methods for recovering verb subcategorization have focused on verb syntax. However, syntax alone is not adequate for applications that require natural language understanding such as machine translation, information retrieval, and data base access. A lexicon for natural language understanding must indicate the possible linkings or mappings of syntactic argument positions onto semantic roles or slots in a frame representation. Semantic role assignment must be sufficient to distinguish subjects and objects with different roles, in order to capture differences between sentences like *We recycle* in which the subject plays the agent role and *This bottle recycles* in which the subject plays the patient role. In addition to subject and object, subcategorization frames must include information about prepositional phrase arguments, in particular, the specific preposition which is the head. This is necessary because a preposition can occur with different semantic roles, depending on the verb as in *The butcher presented John with a knife* where *with a knife* is a patient or theme and *The butcher cut the meat with a knife* where *with a knife* is an instrument.

It is important to notice that when we are dealing with syntax-to-semantics mappings, the number of frames is significantly larger than when we are dealing with syntactic subcategorization alone. For example, Brent ([Bre93]) worked with only seven syntactic frames, while Manning ([Man93]) uses 19, some of which may be parameterized for a preposition. Longman's (LDOCE) lists 31 possible syntactic frames. These frames, while being general enough to cover a large array of data, are not detailed enough for semantic role assignment. An adequate treatment of semantic role assignment must account for all of the possible syntax-to-semantics mappings for each argument position, which involves identifying all of the *surface* configurations that the arguments occur in. Thus active sentences and corresponding passive sentences, dative shifted sentences and the corresponding unshifted sentences, etc. will be treated as separate surface subcategorization frames.[1] B. Levin's ([Lev93]) list of diathesis alternations provides a closer indication of the number of frames that may be required for syntax-to-semantics mappings.

Our preliminary experiments with the Suzanne corpus (see below) confirm our intuition about the large number of distinct frames that are needed to adequately represent syntax-to-semantics mappings. There are, however, many obstacles to acquiring all of the possible mappings for each verb. The obstacles include data sparseness and flexibility in verb class membership.

A large barrier to acquiring complete verb subcategorization is the sparseness of data. This problem is common to all methods that use corpora. According to Zipf's law, there will always be a large tail of data that occurs very infrequently. For example, Church ([Chu88]) states that 40,000 of the words in the Brown corpus occur less than 5 times. Indeed some of data may not occur in the corpus at all. This has prompted researchers to develop many techniques for smoothing probability distributions in order to better approximate the true distribution.

The problem of data sparseness for verb subcategorization is acute. It is very unlikely for a given verb that the set of all subcategorization frames that it can occur in will be represented in a corpus. This would be true even if every verb occurrence could be extracted and used for evidence.[2] Manning ([Man93]), for example, states that many of the more obscure subcategorizations for less common verbs never occurred in his corpus. The problem of data sparseness is even worse if we are looking for syntax-to-semantics mappings, and not just surface subcategorization frames.[3]

In attempting to solve the data sparseness problem, the productivity of most transitivity alternations would seem to work in our favor. An apparent solution is to use verb classes. With a system of verb classification, previously unseen subcategorization frames would be predictable by noting class membership. Class membership would allow us to correctly assign semantic roles in cases of regular polysemies, such as the one involving verbs of sound emission being used as verbs of motion (as in *The bullet whistled through the air* and *The car roared up the driveway*, [Lev93]), even if very few instances of the rule were actually seen during acquisition. Productivity should also work in our favor in predicting possible syntax-to-semantics mappings for newly coined verbs. For example, having seen *I'll modem him tomorrow* we might also predict *I'll modem the news to him* (B. Levin, [Lev93]).

However, according to Dagan et al. ([DPL]), use of a relatively small number of fixed word classes or clusters may cause a substantial loss of information. In the case of verb class membership, the loss of information would stem from the flexibility or vagueness of criteria for class membership. Every definition of class membership (such as what constitutes a change of state verb, a stative verb, a verb of arriving, a verb of attaching, etc.) is bound to fail in cases in which the rules are bent. For example, locative inversion typically applies to stative verbs and verbs of arriving, but examples such as *In this factory work many people* are also found. The input conditions to many transitivity alternations such as the medio passive (*e.g., These books sell well*) are also notoriously hard to characterize. Each exception taken individually may be so rare as to be insignificant, but the aggregate effect of all of the exceptions and rule-bendings is a significant barrier to robustness. As an alternative to fixed classifications, Dagan et al. ([DPL]) argue for the use of a probabilistic similarity-based model, a position that we also advocate. This type of model will allow an NLP system to analyze sentences with previously unseen syntax-to-semantics mappings that are not covered by verb class definitions.

[1] These surface configurations can be handled as alternate linkings of grammatical functions to semantic roles as in LFG or as movements from deep structure positions as in transformational theories of syntax.

[2] No method has yet been able to use all verb occurrences (cf. Manning [Man93]).

[3] The difficulties of dealing with sparse data are compounded by statistical noise and errors in the acquisition program. Data that is not well attested cannot always be trusted. We will not address this particular aspect of the problem in this paper.

Complements		Adjuncts	
s	logical subject	p	place
o	logical direct object	q	direction
S	surface (and not logical) subject	t	time
O	surface (and not logical) direct object	h	manner or degree
i	indirect object	m	modality
u	prepositional object	c	contingency
e	predicate complement of subject	r	respect
j	predicate complement of object	w	comitative
a	agent of passive	k	benefactive
n	particle of phrasal verb	b	absolute
z	complement of catenative		
x	relative clause having higher clause as antecedent		
G	"guest" having no grammatical role within its tagma		

Table 1: Grammatical Functions in the SUSANNE scheme.

Data sparseness and the flexibility of class membership mean that, not only will it be impossible to achieve complete acquisition, but also any natural language system that uses fixed classifications will be inadequate. In the remainder of this paper we will demonstrate the seriousness of the sparseness problem for syntax-to-semantics mappings, and then outline a plan of research for accommodating sparseness and class flexibility in the acquisition of syntax-to-semantics mappings from corpora. The methods that we propose are geared toward making the acquisition of syntax-to-semantics mappings as complete as possible, while at the same time allowing for previously unseen mappings to be analyzed.

2 A Preliminary Experiment Concerning Sparseness

In order to illustrate the extent to which data sparseness is a problem, we ran a preliminary experiment using the SUSANNE Corpus ([Sam92]). The SUSANNE Corpus comprises an approximately 128,000-word subset of the Brown Corpus of American English, annotated in accordance with the SUSANNE scheme. The annotations indicate constituent structure, part of speech, and grammatical relations, among other things.

The purpose of the test was to tabulate subcategorization frames that are tagged in the SUSANNE corpus and determine the percentage of frames in a test set that do not occur in a training set. This should give us an approximate measure of the severity of the sparseness problem. Although the SUSANNE Corpus does not encode semantic role assignments explicitly, it does provide some indication of semantic role. The distinction between logical and non-logical subject and object, for example, refers to the underlying semantic role. In addition, the agent of a passive verb is encoded explicitly. Several of the adjunct functions in the SUSANNE scheme, such as place, direction and benefactive are also related to semantic roles. We will, therefore, consider the frames we extract to approximate syntax-to-semantics mappings, and we will take our results to be an approximation of the severity of the sparseness problem for syntax-to-semantics mappings. Table 1 lists some of the SUSANNE corpus tags for complements and adjuncts ([Sam92]).

To run the test, we divided the SUSANNE corpus into two pieces, the Training set and the Test set. Information about these sets is given in Table 2. An attempt was made to preserve for each set the same proportions of each genre (newspaper, literature, etc.) as existed in the original corpus. For each clause in the training set and test set, we extracted the verb and tags for other phrases occurring in the clause. We will use the term *verb-frame pair* to refer to a verb with a set of tags that occurred in the same clause. Some examples of verb-frame pairs and the sentences that they came from are shown in Table 3. Each unique set of tags was counted as a separate frame, so that each verb typically occurs in several verb-frame pairs, each reflecting a different syntactic context that the verb appeared in.

The results of the experiment are summarized in Table 4. The following information is given. The average number of subcategorization frames per verb in the training set (F/V), the number of unique subcategorization frames in the training set (NF), the percent of verb-frame pairs recognized in the test set after training on the training set (%R), and percent of verb-frame pairs in the test set recognized if only verbs that have been seen before are counted (%RC).

	Training set	Test set
# of words	104,303	23,899
% of corpus	81.4%	18.6%
# of verb occurrences	15460	3532
# of distinct verbs	2240	918

Table 2: Description of data sets.

Verb-Frame Pair	**Sentence**
mean s:N r:P(by) o:D	When I question them as to what they MEAN by concepts like liberty and democracy, . . .
burst s:N o:N	BURSTING paper cartridges, he scattered powder beneath the nearest wagon and . . .
issue t:N r:P(under) S:N	. . . to make the Rural Roads Authority a revolving fund under which new bonds would be ISSUED every time a portion of the old ones are paid off . . .
be s:N p:P(with) b:T	Clayton IS with him, takin him out of the valley.
arm s:N h:P(with)	Ordinary Carey Williams, ARMED with a pistol, stood by at the polls to insure order.
try s:N	I saw you driftin away–but I TRIED.
neighbour s:N	. . . a population and an area appropriate to a pre-World-War-I great power have been, following conquest, rules against their will by a NEIGHBORING people, . . .
hate s:N o:T	"I HATE to leave my garden", Gavin said.
eat s:N o:D	It's not much of a meal, but it's what I EAT.
cast a:P(by) S:N	The two horses broke from the yard, from the circle of light CAST by the lamp still burning in the house, ...
favour a:P(by) S:N	. . . the old reconstructed South–to use the moderate words FAVORED by Mr. Thomas Griffith–finds itself unsympathetic . . .

Table 3: Examples of Verb-Frame Pairs

	Functions Only	Functions & POS	POS Only
Complements Only	F/V: 1.93 NF: 141 %R: 65% %RC: 74%	F/V: 2.20 NF: 339 %R: 61% %RC: 69%	F/V: 2.03 NF: 225 %R: 64% %RC: 73%
Complements & Adjuncts	F/V: 3.90 NF: 2310 %R: 35% %RC: 39%	F/V: 4.27 NF: 3344 %R: 32% %RC: 35%	F/V: 3.91 NF: 2093 %R: 36% %RC: 40%

Table 4: Results of experiment.

Q	quotation
S	main clause
F	finite clause
T	non-finite verbal clause
Z	reduced (”whiz-deleted”) relative clause
L	other verbless clause
A	special “as” clause
W	“with” clause
N	noun phrase
V	verb group
J	adjective phrase
R	adverb phrase
P	prepositional phrase
D	determiner phrase
M	numeral phrase
F	genitive phrase

Table 5: Syntactic Categories in the SUSANNE scheme.

The results are compared with respect to two criteria. The first criterion is whether or not we count adjuncts as well as complements in the subcategorization frames. It may be reasonable to count adjuncts because of possible controversy over what is a complement and what is an adjunct. (See Table 1.) The second criterion is whether we count only grammatical functions or whether we count syntactic categories in addition. Counting syntactic categories means that a frame with a noun phrase direct object will count as different from a frame with a clausal direct object. Table 5 lists the syntactic categories that were counted in our experiment. The list is derived from the list of form tags in the SUSANNE corpus documentation ([Sam92]).

This is of course a very small experiment, but it hints at the severity of the sparseness problem. (Follow-up experiments should work with larger corpora.) The experiment shows two interesting results. First, the number of distinct subcategorization frames (NF) is much larger than the number of frames in LDOCE or the number of frames that Manning and Brent worked with. This is because we are working with *surface* frames, each of which represents a unique syntax-to-semantics mapping. The study of surface frames is important for transformational as well as non-transformational approaches to syntax. A transformational theory that reduces subcategorization classes to a small number of deep structure frames will still have to account for all of the corresponding surface structures eventually. The second interesting result is that 26%–51% (depending on whether adjuncts and parts of speech are counted) of the verb-frame pairs in the test set did not appear in the training set. This means that a lexicon that is acquired based on the training set alone will not provide adequate coverage of unrestricted text. We want to stress that data sparseness is truly a problem that has not yet been dealt with in research on corpus-based acquisition of subcategorization.

3 Ideas for a Solution

The focus of this paper is on showing that data sparseness and flexibility of class membership present problems for acquisition of syntax-to-semantics mappings. Our work on solving these problems is in a very early stage, but we would like to sketch some preliminary ideas for a solution. We will assume that the task is to learn syntax-to-semantics mappings given a text corpus, a parser, and an ontology consisting of concepts and slots with selectional restrictions. A system for acquisition of syntax-to-semantics mappings must include the following components:

- Extraction of syntactic frames from the corpus
- Sense disambiguation of nouns and verbs
- Inferring mapping of syntactic frames onto semantic frames
- Dealing with data sparseness

The first step would be the extraction of syntactic frames from the corpus. This step has been the subject of the most study. However, most methods used for doing this are quite simple. It should be possible to improve on these methods by using more robust methods for locating the verbs and parsing their arguments. Manning ([Man93]), for example, points out several types of structures that cause problems for his method, including coordinate structures and relative clauses. He argues that it is better to generate too much and then use statistics to filter the results. There are, however, cases where his parser recognizes too few subcategorization frames. Better results could be obtained with improved syntactic analysis.

At some point during the acquisition process it will be necessary to provide sense disambiguation for both verbs and their nominal arguments. Disambiguation of verb senses will provide a more accurate basis for similarity clustering (see below). Noun senses need to be identified for two purposes—to help with similarity clustering of verbs based on the semantics of the nouns that they occur with and to establish mappings from syntactic argument positions to semantic slots based on semantic selectional restrictions.

The next step in acquiring syntax-to-semantics mappings will be to determine the mapping between syntactic arguments and semantic roles as given by a core lexicon or ontology. The mapping will be established by matching noun senses against selectional restrictions on semantic slots in the ontology. If the nouns filling a syntactic slot meet the selectional restrictions on a semantic slot, a mapping will be established between the syntactic and semantic slots. This procedure will result in both overgeneration and undergeneration of mappings. Overgeneration of mappings will occur in cases that involve multiple arguments with roughly the same selectional restrictions. In order to avoid hypothesizing multiple mappings, it may be possible in some cases to apply some heuristics concerning typical mappings such as the association of agent with subject. In contrast, the unreliability of hand coded selectional restrictions will result in undergeneration of mappings because noun phrases will often not meet any of the verb's selectional restrictions. If the procedure for association of phrases with ontological slots can be constrained enough to avoid overgeneration of frames and extended enough to allow for flexibility in application of selectional restrictions, it should be possible to hypothesize syntax-to-semantics mappings with having to resort to user interaction. Liu and Soo ([LS93]) use a similar method to tackle the harder problem of acquiring semantic roles from scratch. However, they rely heavily on the user, asking 113 questions for every 100 semantic roles.

Finally, we will have to deal with data sparseness. Our goal will be to perform correct syntax-to-semantics mappings for verbs in previously unattested subcategorization frames. There are at least three ways of accommodating sparseness, as outlined by Dagan et al. ([DMM93]): smoothing methods, class based methods, and similarity methods. While both class based methods and similarity methods have advantages over simple smoothing techniques, Dagan et al. argue that unrestricted language use may not be adequately captured by class based methods. We believe that the similarity approach is currently the best solution to data sparseness. We are currently experimenting with a measure of verb similarity based on the syntactic frames that the verbs occur in and the semantic characterization of their complements obtained from WordNet. The similarity measure will be used in assigning meanings to sentences with verbs in subcategorization frames that they have not previously been seen in. The assignment of meaning will be based on the syntax-to-semantics mappings of similar verbs that have been observed in the same frame.

4 Conclusion

We have shown that data sparseness can be a major obstacle to the acquisition of syntax-to-semantics mappings from corpora. A lexicon that is acquired based on a training corpus might not have adequate coverage of unrestricted text. Grouping verbs into classes is only a partial solution to this problem because flexibility in the criteria for class membership results in unexpected subcategorization frames. We believe that what is needed in place of hard-and-fast verb classes is a statistical notion of similarity among verbs. Our future research will be concerned with identifying an adequate similarity metric.

Acknowledgements: We would like to thank Chris Manning for helpful comments on an earlier draft of this paper.

References

[Bre93] Michael R. Brent. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*, 19(2):243–262, June 1993.

[Chu88] Kenneth W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings, Second Conference on Applied Natural Language Processing*, pages 136–143. ACL, 1988.

[CM93] Kenneth W. Church and Robert L. Mercer. Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1):1–24, March 1993.

[DMM93] Ido Dagan, Shaul Markus, and Shaul Markovitch. Contextual word similarity and estimation from sparse data. In *Proceedings of the 30th Annual Meeting of the ACL*, pages 164–171, Morristown, NJ, 1993. Association for Computational Linguistics.

[DPL] Ido Dagan, Fernando Pereira, and Lillian Lee. Similarity-based estimation of word cooccurrence probabilities. To appear in the Proceedings of the 32nd Annual Meeting of the ACL, New Mexico State University, June 1994.

[Lev93] Beth Levin. *English Verb Classes and Alternations.* The University of Chicago Press, Chicago and London, 1993.

[LS93] Rey-Long Liu and Von-Wun Soo. An empirical study on thematic knowledge acquisition based on syntactic clues and heuristics. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 243–250, 1993.

[Man93] Christopher D. Manning. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the ACL*, pages 235–242, 1993.

[Sam92] Geoffrey Sampson. *The Susanne Corpus.* School of Cognitive & Computing Sciences, University of Sussex, 1 edition, September 1992.