Probabilistic Models of Verb-Argument Structure

Daniel Gildea
Dept. of Computer and Information Science
University of Pennsylvania
dgildea@cis.upenn.edu

Abstract

We evaluate probabilistic models of verb argument structure trained on a corpus of verbs and their syntactic arguments. Models designed to represent patterns of verb alternation behavior are compared with generic clustering models in terms of the perplexity assigned to held-out test data. While the specialized models of alternation do not perform as well, closer examination reveals alternation behavior represented implicitly in the generic models.

1 Introduction

Recent research into verb-argument structure has has attempted to acquire the syntactic alternation behavior of verbs directly from large corpora. McCarthy (2000), Merlo and Stevenson (2001), and Schulte im Walde (2000) have evaluated their systems' accuracy against human judgments of verb classification, with the comprehensive verb classes of Levin (1993) often serving as a gold standard. Another area of research has focused on automatic clustering algorithms for verbs and their arguments with the goal of finding groups of semantically related words (Pereira et al., 1993; Rooth et al., 1999), without focusing specifically on alternation behavior. We aim to bring these strands of research together with a unified probabilistic model of verb argument structure incorporating alternation behavior.

Unraveling the mapping between syntactic functions such as subject and object and semantic roles such as agent and patient is an important piece of the language understanding problem. Learning the alternation behavior of verbs automatically from unannotated text would significantly reduce the amount of labor needed to create text understanding systems, whether that labor takes the form of writing lexical entries or of annotating semantic information to train statistical systems.

Our use of generative probabilistic models of argument structure also allows for language modeling applications independent of semantic interpretation. Language models based on head-modifier lexical dependencies in syntactic trees have been shown to have lower perplexity than n-gram language models and to reduce word-error rates for speech recognition (Chelba and Jelinek, 1999; Roark, 2001). Incorporating semantic classes and verb alternation behavior could improve such models' performance. Automatically derived word clusters are used in the statistical parsers of Charniak (1997) and Magerman (1995). Incorporating alternation behavior into such models might improve parsing results as well.

This paper focuses on evaluating probabilistic models of verb-argument structure in terms of how well they model unseen test data, as measured by perplexity. We will examine maximum likelihood bigram and trigram models, clustering models based on those of Rooth et al. (1999), as well as a new probabilistic model designed to capture alternations in verb-argument structure.

2 Capturing Alternation Behavior

Automatic clustering of co-occurrences of verbs and their direct objects was first used to induce semantically related classes of both verbs and nouns (Pereira et al., 1993). Rooth et al. (1999) used the Expectation Maximization algorithm to perform soft clustering by optimizing the parameters of a fairly simple probability model, which considers the verb and noun to be independent given the unobserved cluster variable c:

$$P(v, n) = \sum_c P(c)P(v|c)P(n|c)$$

In Rooth et al. (1999), the variable v represented not only the lexical verb but also its syntactic relation to the noun: either direct object, subject of an intransitive, or subject of a transitive verb.

However, the relationship between the underlying, semantic arguments of a verb and the syntactic roles in a sentence is not always straightforward. Many verbs exhibit *alternations* in their syntactic behavior, as shown by the following examples:

(1) The Federal Reserve increased rates by 1/4%.

(2) Interest rates have increased sharply over the past year.

The noun *rates* appears as the syntactic object of the verb *increase* in the first sentence, but as its subject in the second sentence, where the verb is used intransitively, that is, without an object. One of the clusters found by the model of Rooth et al. (1999) corresponded to "verb of scalar change" such as *increase*, *rise*, and *decrease*. The model places both subject-of-intransitive-*increase* and direct-object-of-*increase* in this class, but does not explicitly capture the fact that these to values represent different uses of the same verb.

The phenomenon of verb argument alternations has been most comprehensively studied by Levin (1993), who catalogs over 3,000 verbs into classes according to which alternations they participate in. A central thesis of Levin's work is that a verb's syntactic alternations are related to its semantics, and that semantically related verb will share the same alternations. For example, the alternation of examples 1 and 2 is shared by verbs such as *decrease* and *diminish*.

Table 1 gives the most common nouns occurring as arguments of selected verbs in our corpus, showing how alternation behavior shows up in corpus statistics. The verbs *open* and *increase*, classified by Levin and others as exhibiting a causative alternation between transitive and intransitive usages, share many of the same nouns in direct object and subject-of-intransitive positions, as we would expect. For example, *number*, *cost*, and *rate* occur among the ten most common nouns in both positions for *increase*, and themselves seem semantically related. For *open*, the first three words in either position are the same. For the verb *play*, on the other hand, classified as an "object-drop" verb by Merlo and Stevenson (2001), we would expect overlap between the subject of transitive and intransitive uses. This is in fact the case, with *child*, *band*, and *team* appearing among the top ten nouns for both positions. However, *play* also exhibits an alternation between the direct object and subject of intransitive positions for *music*, *role*, and *game*. These two sets of nouns seem to fill different semantic roles of the verb, the first set being agents and the second being themes. This example illustrate the complex interaction between verb sense and alternation behavior: "The band played" and the "The music played" are considered to belong to different senses of play by WordNet (Fellbaum, 1998) and other word sense inventories. However, it is interesting to note that nouns from both the broad senses of *play*, "play a game" and "play music", participate in both alternations. An advantage of our EM-based soft clustering algorithm is that it can assign a verb to multiple clusters; ideally, we would hope that a verb's clusters would correspond to its senses.

We expect verbs which take similar sets of argument fillers to be semantically related, and to participate in the same alternations. This idea has been used by McCarthy (2000) to identify verbs participating in specific alternations by looking for overlap between nouns used in different positions, and by using WordNet to classify role fillers into semantic categories. Schulte im Walde (2000) uses an EM-based automatic clustering of verbs to attempt to derive Levin classes from unlabeled data. As in McCarthy (2000), the nouns are classified using WordNet. However, the appearance of the same noun in different syntactic positions is not explicitly captured by the probability model used for clustering.

This observation motivated a new probabilistic model of verb argument structure designed to explicitly capture alternation behavior. In addition to an unobserved cluster variable c, we introduce a second unobserved variable r for the semantic role of an argument. The role r is dependent on both the cluster c to which our verb-noun pair belongs, and the syntactic slot s in which the noun is found, and the probability of an observed triple $P(v, s, n)$ is estimated as:

$$\sum_{c,r} P(c)P(v|c)P(s|c)P(r|c,s)P(n|r,c)$$

The noun is independent of the verb given the cluster variable, as before, and the noun is independent of the syntactic slot s given the cluster c and the semantic role r. The semantic role variable r can take two values, with $P(r|c,s)$ representing the mapping from syntax to semantic role for a cluster of verbs. We expect the clusters to consist of verbs that not only appear with the same set of nouns, but share the same mapping from syntactic position to semantic role. For example *increase* and *decrease* might belong to same cluster as they both appear frequently

Verb	*Object*	*Subj of Intransitive*	*Subj of Transitive*
close	door	door	troop
	eyes	eyes	door
	mouth	mouth	police
	firebreak	exhibition	gunman
	way	shop	woman
	possibility	show	man
	gate	trial	guard
	account	conference	soldier
	window	window	one
	shop	gate	company
increase	risk	number	government
	number	proportion	increase
	share	population	use
	profit	rate	effect
	lead	pressure	sale
	pressure	amount	level
	rate	cost	presence
	likelihood	sale	Party
	chance	rates	Labour
	cost	profit	bank
play	part	child	band
	role	band	factor
	game	team	England
	host	role	child
	music	player	people
	card	game	woman
	piano	smile	man
	tennis	people	team
	parts	music	all
	guitar	boy	group

Table 1: Examples from the corpus: most common arguments for selected verbs

with *rate*, *number*, and *price* in both the direct object and subject of intransitive slots, and would assign the same value of r to both positions. The verb *lower* might belong to a different cluster because, although it appears with the same nouns, they appear as the direct object but not as the subject.

The Expectation Maximization algorithm is used to train the model from the corpus, iterating over an Expectation step in which expected values for the two unobserved variables c and r are calculated for each observation in the training data, and a Maximization step in which the parameter of each of the five distributions $P(c)$, $P(v|c)$, $P(s|c)$, $P(r|c,s)$, and $P(n|n,c)$ are set to maximize the likelihood of the data given the expectations for c and r.

3 The Data

For our experiments we used a version of the British National Corpus parsed with the statistical parser of Collins (1997). Subject and direct object relations were extracted by searching for NP nodes dominated by S and VP nodes respectively. The head words of the resulting subject and object nodes were found using the deterministic headword rules employed by the parsing model. The individual observations of our dataset are noun-verb pairs of three types: direct object, subject of a verb with an object, and subject of a verb without an object. As a result, the subject and object relations of the same original sentence are considered independently by all of the models we examine.

Direct object noun phrases were assigned the function tags of the Treebank-2 annotation style (Marcus et al., 1994) in order to distinguish noun phrases such as temporal adjuncts from true direct objects. For example, in the sentence "He ate yesterday", *yesterday* would be assigned the Temporal tag, and therefore not considered a direct object for our purposes. Similarly, in the sentence "Interest rates rose 2%", *2%* would be assigned the Extent tag, and this instance of *rise* would be considered intransitive.

Function tags were assigned using a simple probability model trained on the Wall Street Journal data from the Penn Treebank, in a technique similar to that of Blaheta and Charniak (2000). The model predicts the function tag conditioned on the verb and head noun of the noun phrase:

$$P(f|v,n) = \begin{cases} \tilde{P}(f|v,n) & (v,n) \in T \\ \frac{1}{2}\tilde{P}(f|v) + \frac{1}{2}\tilde{P}(f|n) & \text{otherwise} \end{cases}$$

where f ranges over the function tags defined (Marcus et al., 1994), or the null tag. Only cases assigned the null tag by this model were considered true direct objects. Evaluated on the binary task of whether to assign a function tag to noun phrases in object position, this classifier was correct 95% of the time on held-out data from the Wall Street Journal. By never assigning a function tag, one would achieve 85% accuracy. While we have no way to evaluate its accuracy on the British National Corpus, certain systematic errors are apparent. For example, while it classifies *2%* as an Extent in "Interest rates increased 2%", it assigns no tag to *crack* in "The door opened a crack". This type of error leads to the appearance of *door* as a subject on transitive uses of *open* in Table 1.

Both verbs and nouns were lemmatized using the XTAG morphological dictionary (XTAG Research Group, 2001). As we wished to focus on alternation behavior, verbs that were used intransitively than 90% of the time were excluded from the data; we envision that they would be handled by a separate probability model. Pronouns were excluded from the dataset, as were verbs and nouns that occurred fewer than 10 times, resulting in a vocabulary of 4,456 verbs and 17,345 nouns. The resulting dataset consisted of 1,372,111 triples of verb, noun, and syntactic relation. Of these, 90% were used as training material, 5% were used as a cross-validation set for setting linear interpolation and deterministic annealing parameters, and 5% were used as test data for the results reported below.

4 The Models

We compare performance of a number of probability models for our verb argument data in order to explore the dependencies of the data and the impact of clustering. Graphical representations of the clustering models are shown in Figure 1.

Unigram Baseline: This model assumes complete independence of the verb, syntactic slot, and noun, and serves to provide a baseline for the complexity of the task:

$$P_1(v,s,n) = P(v)P(s)P(n)$$

Bigram: This model predicts both the noun and syntactic slot conditioned on the verb, but independently of one another:

$$P_2(v,s,n) = P(v)P(s|v)P(n|v)$$

Trigram: This is simply the empirical distribution over triples of verb, slot, and noun:

$$P_3(v,s,n) = \lambda P(v,s,n)$$

Three-way Aspect: Following Hofmann and Puzicha (1998), we refer to EM-based clustering as the aspect model, where different values of the cluster variable are intended to represent abstract "aspects" of the data. The simplest version of the clustering model predicts verb, slot, and noun independently given the cluster variable c:

$$P_c(v,s,n) = P(c)P(v|c)P(s|c)P(n|c)$$

with all four component distributions being estimated by EM training.

Verb-Slot Aspect: This is the model of Rooth et al. (1999), in which the verb and slot are combined into one atomic variable before the aspect model is trained:

$$P_{c_{vs}} = P(c)P(v,s|c)P(n|c)$$

Noun-Slot Aspect: A variation on the above model combines the slot with the noun, rather than the verb:

$$P_{c_{ns}} = P(c)P(v|c)P(n,s|c)$$

Alternation: This model, described in more detail above, introduces a new unobserved variable r for the semantic role of the noun, which can take two values:

$$P_{alt} = P(c)P(v|c)P(s|c)P(r|s,c)P(n|r,c)$$

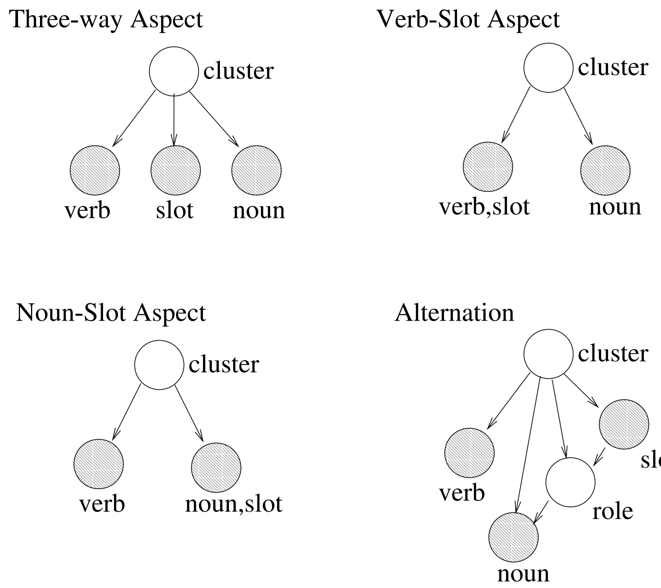

Figure 1: Graphical models: shading represents observed variables, arrows probabilistic dependencies.

Fixed Alternation: This model is designed to incorporate the assumption that the semantic roles of the subject and object of the same verb must be different. The independence assumptions are identical to those of the simple alternation model:

$$P_{alt_2} = P(c)P(v|c)P(s|c)P(r|s,c)P(n|r,c)$$

but the probability $P(r|s,c)$ is only trained for $s =$ subj-intrans. The model is constrained to assign one value of the role variable to direct objects, $P(r = 0|s = \text{obj}) = 1$ and the other role to subjects of transitives: $P(r = 1|s = \text{subj-trans}) = 1$.

5 Results

Perplexity results on held-out test data for each of the models are shown in Table 2. Because models 2, 3, 5, and 6 will assign zero probability to certain pairs of values not seen in the training data, they were combined with the unigram baseline model in order to obtain a perplexity over the entire test set comparable to the other models. This was done using linear interpolation, with the interpolation weight optimized on the cross-validation data. Perplexity is the geometric mean of the reciprocal of the probability assigned by the model to each triple of verb, noun, and slot in the test data:

$$PP = e^{-\frac{1}{N}\sum_i \log P(v_i,n_i,s_i)}$$

For the single-variable clustering models (4, 5 and 6) 128 values were allowed for the cluster variable c. For the two-variable clustering models (7 and 8), 64 values for c and 2 values for the unobserved semantic roles variable r were used, making for a total of 128 distributions over nouns ($P(n|r,c)$) but only 64 over verbs ($P(v|c)$). The total number of parameters for each model is shown in Table 2. Because deterministic annealing was used to smooth the probability distributions for each cluster and prevent overfitting the training data, the perplexities obtained were relatively insensitive to the number of clusters used.

Of the clustering models, the Verb-Slot Aspect model did the best, with a perplexity of 2.31M. It is perhaps surprising how close the Three-way Aspect model came, with a perplexity of 2.41M, despite the fact that it models the noun as being independent of the syntactic position for a given verb. One explanation for this is that nouns in fact occur in all three positions more frequently than we would expect from traditional accounts of alternation behavior. This is shown in our corpus examples of Table 1 by the high frequency of *door* as a subject of an transitive use of *open*. Even in the traditional alternation pattern where a noun occurs in two of the three positions, the Three-way Aspect model may do better at capturing this overlap, even though it will mistakenly assign probability mass to the same nouns appearing in the third syntactic position, than do models 5 and 6, which are not able to generalize

Model	*Test Perplexity*	*Total Parameters*
1. Unigram Baseline	5.50M	20,651
2. Bigram	2.95M	57.64M
3. Trigram	2.55M	172.88M
4. Three-way Aspect	2.41M	2.64M
5. Verb-Slot Aspect	2.31M	3.47M
6. Noun-Slot Aspect	2.66M	6.56M
7. Alternation	2.57M	2.43M
8. Fixed Alternation	2.60M	2.43M
9. Trigram+Verb-Slot Aspect	2.06M	176.36M

Table 2: Comparison of probability models

at all across the different arguments of a given verb.

The models specifically designed to capture alternation behavior (7 and 8) did not do as well as the generic clustering models. One explanation is that the unconstrained models are able to fit the data better by clustering together specific arguments of different verbs even when the two verbs do not share the same alternation behavior. Examining the clusters found by the Verb-Slot Aspect shows that it in fact seems to find alternation behavior for specific verbs despite the model's inability to explicitly represent alternation. In many cases, two roles of the same verb are assigned to the same cluster. Examples of the top ten members of sample clusters are shown in Table 3. Examining the sample verbs of Table 1, we see that the model assigns the direct object and subject of intransitive slots of *open* to the same cluster, implicitly representing the verb's alternation behavior, and in fact does the same for the semantically related verbs *close* and *shut*. Similarly, the direct object and subject of intransitive slots of *increase* are assigned to the same cluster. However, in an example of how the model can cluster semantically related verbs that do not share the same alternation behavior, the direct object slot of *reduce* and the subject of transitive slot of *exceed* are groups together with *increase*. Of particular interest is the verb *play*, for which the model assigns one cluster to each of the alternation patterns noted in Table 1. Cluster 18 represents the alternation between direct object and subject of intransitive seen with *part*, *game*, and *music*, while cluster 92 represents the agent relation expressed by subjects of both transitive and intransitive sentences.

The final line of Table 2 represents an interpolation of the best n-gram and best clustering model, which further reduces perplexity to 2.06 million.

6 Conclusion

We have attempted to learn the mapping from syntactic position to semantic role in an unsupervised manner, and have evaluated the results in terms of our systems' success as language model for unseen data. The models designed to explicit represent verb alternation behavior did not perform as well by this metric as other, simpler probability models.

A perspective on this work can be gained by comparison with attempts at unsupervised learning of other natural language phenomena including part-of-speech tagging (Merialdo, 1994) and syntactic dependencies (Carroll and Charniak, 1992; Paskin, 2001). While models trained using the Expectation Maximization algorithm do well at fitting the data, the results may not correspond to the human analyses they were intended to learn. Language does not exist in the abstract, but conveys information about the world, and the ultimate goal of grammar induction is not just to model strings but to extract this information. This suggests that although the probability models constrained to represent verb alternation behavior did not achieve the best perplexity results, they may be useful as part of an understanding system which assigns semantic roles to arguments. The implicit representation of alternation behavior in our generic clustering model also suggests using its clusters to initialize a more complex model capable of assigning semantic roles.

Acknowledgments This work was undertaken with funding from the Institute for Research in Cognitive Science at the University of Pennsylvania and DoD Grant MDA904-00C-2136.

References

Don Blaheta and Eugene Charniak. 2000. Assigning function tags to parsed text. In *Proceedings of the 1st Annual Meeting of the North American Chapter of the ACL (NAACL)*, pages 234–240, Seattle, Washington.

Cluster Id	*Noun*	*Verb-Slot*	*Cluster Id*	*Noun*	*Verb-Slot*
57	door	open-obj	18	part	play-obj
	mouth	open-subj-intrans		role	form-obj
	eyes	close-obj		lip	take-obj
	firebreak	close-subj-intrans		game	bite-obj
	gate	shut-subj-intrans		basis	play-subj-intrans
	shop	slam-subj-intrans		host	lick-obj
	window	shut-obj		parts	curl-subj-intrans
	way	knock-obj		music	see-obj
	exhibition	reach-obj		card	constitute-obj
47	number	increase-subj-intrans	92	people	play-subj-intrans
	amount	require-obj		man	win-subj-intrans
	supply	reduce-obj		child	take-subj-trans
	level	increase-obj		woman	make-subj-trans
	rate	exceed-subj-trans		one	need-subj-trans
	tooth	need-obj		the	play-subj-trans
	income	include-obj		band	see-subj-trans
	risk	affect-obj		group	get-subj-intrans
	activity	show-obj		team	manage-subj-intrans

Table 3: Sample Clusters from Verb-Slot Aspect Model

Glenn Carroll and Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. In *Workshop Notes for Statistically-Based NLP Techniques*, pages 1–13. AAAI.

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *AAAI-97*, pages 598–603, Menlo Park, August. AAAI Press.

Ciprian Chelba and Frederick Jelinek. 1999. Recognition performance of a structured language model. In *EUROSPEECH*.

Michael Collins. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th ACL*, pages 16–23, Madrid, Spain.

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

Thomas Hofmann and Jan Puzicha. 1998. Statistical models for co-occurrence data. Memo, Massachussetts Institute of Technology Artificial Intelligence Laboratory, February.

Beth Levin. 1993. *English Verb Classes And Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

David Magerman. 1995. Statistical decision-tree models for parsing. In *Proceedings of the 33rd ACL*, Cambridge, Massachusetts.

Mitchell P. Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating predicate argument structure. In *ARPA Human Language Technology Workshop*, pages 114–119, Plainsboro, NJ. Morgan Kaufmann.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *Proceedings of the 1st NAACL*, pages 256–263, Seattle, Washington.

Bernard Merialdo. 1994. Tagging English text with a probabilistic model. *Computational Linguistics*, 20(2):155–172.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distribution of argument structure. *Computational Linguistics*, 27(3), September.

Mark Paskin. 2001. Grammatical bigrams. In T. Dietterich, S. Becker, and Z. Gharahmani, editors, *Advances in Neural Information Processing Systems (NIPS) 14*. MIT Press.

Fernando Pereira, Naftali Tishby, and Lillian Lee. 1993. Distributional clustering of English words. In *Proceedings of the 31st ACL*, pages 183–190, Columbus, Ohio. ACL.

Brian Roark. 2001. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27(2):249–276.

Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. 1999. Inducing a semantically annotated lexicon via EM-based clustering. In *Proceedings of the 37th Annual Meeting of the ACL*, pages 104–111, College Park, Maryland.

Sabine Schulte im Walde. 2000. Clustering verbs semantically according to their alternation behaviour. In *In Proceedings of the 18th International Conference on Computational Linguistics (COLING-00)*, pages 747–753, Saarbrücken, Germany.

XTAG Research Group. 2001. A lexicalized tree adjoining grammar for English. Technical Report IRCS-01-03, IRCS, University of Pennsylvania.