Learning Syntactic Verb Frames Using Graphical Models

Thomas Lippincott
University of Cambridge
Computer Laboratory
United Kingdom
tl318@cam.ac.uk

Diarmuid Ó Séaghdha
University of Cambridge
Computer Laboratory
United Kingdom
do242@cam.ac.uk

Anna Korhonen
University of Cambridge
Computer Laboratory
United Kingdom
alk23@cam.ac.uk

Abstract

We present a novel approach for building verb subcategorization lexicons using a simple graphical model. In contrast to previous methods, we show how the model can be trained without parsed input or a predefined subcategorization frame inventory. Our method outperforms the state-of-the-art on a verb clustering task, and is easily trained on arbitrary domains. This quantitative evaluation is complemented by a qualitative discussion of verbs and their frames. We discuss the advantages of graphical models for this task, in particular the ease of integrating semantic information about verbs and arguments in a principled fashion. We conclude with future work to augment the approach.

1 Introduction

Subcategorization frames (SCFs) give a compact description of a verb's syntactic preferences. These two sentences have the same sequence of lexical syntactic categories (VP-NP-SCOMP), but the first is a simple transitive ("X understood Y"), while the second is a ditransitive with a sentential complement ("X persuaded Y that Z"):

1. Kim (VP understood (NP the evidence (SCOMP that Sandy was present)))

2. Kim (VP persuaded (NP the judge) (SCOMP that Sandy was present))

An SCF lexicon would indicate that "persuade" is likely to take a direct object and sentential complement (NP-SCOMP), while "understand" is more likely to take just a direct object (NP). A comprehensive lexicon would also include semantic information about selectional preferences (or restrictions) on argument heads of verbs, diathesis alternations (i.e. semantically-motivated alternations between pairs of SCFs) and a mapping from surface frames to the underlying predicate-argument structure. Information about verb subcategorization is useful for tasks like information extraction (Cohen and Hunter, 2006; Rupp et al., 2010), verb clustering (Korhonen et al., 2006b; Merlo and Stevenson, 2001) and parsing (Carroll et al., 1998). In general, tasks that depend on predicate-argument structure can benefit from a high-quality SCF lexicon (Surdeanu et al., 2003).

Large, manually-constructed SCF lexicons mostly target general language (Boguraev and Briscoe, 1987; Grishman et al., 1994). However, in many domains verbs exhibit different syntactic behavior (Roland and Jurafsky, 1998; Lippincott et al., 2010). For example, the verb "develop" has specific usages in newswire, biomedicine and engineering that dramatically change its probability distribution over SCFs. In a few domains like biomedicine, the need for focused SCF lexicons has led to manually-built resources (Bodenreider, 2004). Such resources, however, are costly, prone to human error, and in domains where new lexical and syntactic constructs are frequently coined, quickly become obsolete (Cohen and Hunter, 2006). Data-driven methods for SCF acquisition can alleviate

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 420–429,
Jeju, Republic of Korea, 8-14 July 2012.

these problems by building lexicons tailored to new domains with less manual effort, and higher coverage and scalability.

Unfortunately, high quality SCF lexicons are difficult to build automatically. The argument-adjunct distinction is challenging even for humans, many SCFs have no reliable cues in data, and some SCFs (e.g. those involving control such as type raising) rely on semantic distinctions. As SCFs follow a Zipfian distribution (Korhonen et al., 2000), many genuine frames are also low in frequency. State-of-the-art methods for building data-driven SCF lexicons typically rely on parsed input (see section 2). However, the treebanks necessary for training a high-accuracy parsing model are expensive to build for new domains. Moreover, while parsing may aid the detection of some frames, many experiments have also reported SCF errors due to noise from parsing (Korhonen et al., 2006a; Preiss et al., 2007).

Finally, many SCF acquisition methods operate with predefined SCF inventories. This subscribes to a single (often language or domain-specific) interpretation of subcategorization *a priori*, and ignores the ongoing debate on how this interpretation should be tailored to new domains and applications, such as the more prominent role of adjuncts in information extraction (Cohen and Hunter, 2006).

In this paper, we describe and evaluate a novel probabilistic data-driven method for SCF acquisition aimed at addressing some of the problems with current approaches. In our model, a Bayesian network describes how verbs choose their arguments in terms of a small number of frames, which are represented as distributions over syntactic relationships. First, we show that by allowing the inference process to automatically define a probabilistic SCF inventory, we outperform systems with hand-crafted rules and inventories, using identical syntactic features. Second, by replacing the syntactic features with an approximation based on POS tags, we achieve state-of-the-art performance without relying on error-prone unlexicalized or domain-specific lexicalized parsers. Third, we highlight a key advantage of our method compared to previous approaches: the ease of integrating and performing joint inference of additional syntactic and semantic information. We describe how we plan to exploit this in our future research.

2 Previous work

Many state-of-the-art SCF acquisition systems take *grammatical relations* (GRs) as input. GRs express binary dependencies between lexical items, and many parsers produce them as output, with some variation in inventory (Briscoe et al., 2006; De Marneffe et al., 2006). For example, a subject-relation like "ncsubj(HEAD, DEPENDENT)" expresses the fact that the lexical item referred to by HEAD (such as a present-tense verb) has the lexical item referred to by DEPENDENT as its subject (such as a singular noun). GR inventories include direct and indirect objects, complements, conjunctions, among other relations. The dependency relationships included in GRs correspond closely to the head-complement structure of SCFs, which is why they are the natural choice for SCF acquisition.

There are several SCF lexicons for general language, such as ANLT (Boguraev and Briscoe, 1987) and COMLEX (Grishman et al., 1994), that depend on manual work. VALEX (Preiss et al., 2007) provides SCF distributions for 6,397 verbs acquired from a parsed general language corpus via a system that relies on hand-crafted rules. There are also resources which provide information about both syntactic and semantic properties of verbs: VerbNet (Kipper et al., 2008) draws on several hand-built and semi-automatic sources to link the syntax and semantics of 5,726 verbs. FrameNet (Baker et al., 1998) provides semantic frames and annotated example sentences for 4,186 verbs. PropBank (Palmer et al., 2005) is a corpus where each verb is annotated for its arguments and their semantic roles, covering a total of 4,592 verbs.

There are many language-specific SCF acquisition systems, e.g. for French (Messiant, 2008), Italian (Lenci et al., 2008), Turkish (Han et al., 2008) and Chinese (Han et al., 2008). These typically rely on language-specific knowledge, either directly through heuristics, or indirectly through parsing models trained on treebanks. Furthermore, some require labeled training instances for supervised (Uzun et al., 2008) or semi-supervised (Han et al., 2008) learning algorithms.

Two state-of-the-art data-driven systems for English verbs are those that produced VALEX, Preiss et al. (2007), and the BioLexicon (Venturi et al., 2009).

The Preiss system extracts a verb instance's GRs using the Rasp general-language unlexicalized parser (Briscoe et al., 2006) as input, and based on handcrafted rules, maps verb instances to a predefined inventory of 168 SCFs. Filtering is then performed to remove noisy frames, with methods ranging from a simple single threshold to SCF-specific hypothesis tests based on external verb classes and SCF inventories. The BioLexicon system extracts each verb instance's GRs using the lexicalized Enju parser tuned to the biomedical domain (Miyao, 2005). Each unique GR-set considered a potential SCF, and an experimentally-determined threshold is used to filter low-frequency SCFs.

Note that both methods require extensive manual work: the Preiss system involves the *a priori* definition of the SCF inventory, careful construction of matching rules, and an unlexicalized parsing model. The BioLexicon system induces its SCF inventory automatically, but requires a lexicalized parsing model, rendering it more sensitive to domain variation. Both rely on a filtering stage that depends on external resources and/or gold standards to select top-performing thresholds. Our method, by contrast, does not use a predefined SCF inventory, and can perform well without parsed input.

Graphical models have been increasingly popular for a variety of tasks such as distributional semantics (Blei et al., 2003) and unsupervised POS tagging (Finkel et al., 2007), and sampling methods allow efficient estimation of full joint distributions (Neal, 1993). The potential for joint inference of complementary information, such as syntactic verb and semantic argument classes, has a clear and interpretable way forward, in contrast to the pipelined methods described above. This was demonstrated in Andrew et al. (2004), where a Bayesian model was used to jointly induce syntactic and semantic classes for verbs, although that study relied on manually annotated data and a predefined SCF inventory and MLE. More recently, Abend and Rappoport (2010) trained ensemble classifiers to perform argument-adjunct disambiguation of PP complements, a task closely related to SCF acquisition. Their study employed unsupervised POS tagging and parsing, and measures of selectional preference and argument structure as complementary features for the classifier.

Finally, our task-based evaluation, verb clustering with Levin (1993)'s alternation classes as the gold standard, was previously conducted by Joanis and Stevenson (2003), Korhonen et al. (2008) and Sun and Korhonen (2009).

3 Methodology

In this section we describe the basic components of our study: feature sets, graphical model, inference, and evaluation.

3.1 Input and feature sets

We tested several feature sets either based on, or approximating, the concept of *grammatical relation* described in section 2. Our method is agnostic regarding the exact definition of GR, and for example could use the Stanford inventory (De Marneffe et al., 2006) or even an entirely different lexico-syntactic formalism like CCG supertags (Curran et al., 2007). In this paper, we distinguish "true GRs" (tGRs), produced by a parser, and "pseudo GRs" (pGRs), a POS-based approximation, and employ subscripts to further specify the variations described below. Our input has been parsed into Rasp-style tGRs (Briscoe et al., 2006), which facilitates comparison with previous work based on the same data set.

We'll use a simple example sentence to illustrate how our feature sets are extracted from CONLL-formatted data (Nivre et al., 2007). The CONLL format is a common language for comparing output from dependency parsers: each lexical item has an index, lemma, POS tag, tGR in which it is the dependent, and index to the corresponding head. Table 1 shows the relevant fields for the sentence "We run training programmes in Romania and other countries".

We define the feature set for a verb occurrence as the counts of each GR the verb participates in. Table 2 shows the three variations we tested: the simple tGR type, with parameterization for the POS tags of head and dependent, and with closed-class POS tags (determiners, pronouns and prepositions) lexicalized. In addition, we tested the effect of limiting the features to subject, object and complement tGRs, indicated by adding the subscript "lim", for a total of six tGR-based feature sets.

While ideally tGRs would give full informa-

Index	Lemma	POS	Head	tGR
1	we	PPIS2	2	ncsubj
2	run	VV0	0	_
3	training	NN1	4	ncmod
4	programme	NN2	2	dobj
5	in	II	4	ncmod
6	romania	NP1	7	conj
7	and	CC	5	dobj
8	other	JB	9	ncmod
9	country	NN2	7	conj

Table 1: Simplified CONLL format for example sentence "We run training programmes in Romania and other countries". Head=0 indicates the token is the root.

Name	Features	
tGR	ncsubj	dobj
tGR_{param}	ncsubj(VV0,PPIS2)	dobj(VV0,NN2)
$tGR_{param,lex}$	ncsubj(VV0,PPIS2-we)	dobj(VV0,NN2)

Table 2: True-GR features for example sentence: note there are also $tGR_{*,lim}$ versions of each that only consider subjects, objects and complements and are not shown.

tion about the verb's syntactic relationship to other words, in practice parsers make (possibly premature) decisions, such as deciding that "in" modifies "programme", and not "run" in our example sentence. An unlexicalized parser cannot distinguish these based just on POS tags, while a lexicalized parser requires a large treebank. We therefore define *pseudo-GRs* (pGRs), which consider each (distance, POS) pair within a given window of the verb to be a potential tGR. Table 3 shows the pGR features for the test sentence using a window of three. As with tGRs, the closed-class tags can be lexicalized, but there are no corresponding feature sets for *param* (since they are already built from POS tags) or *lim* (since there is no similar rule-based approach).

Name	Features			
pGR	-1(PPIS2)	1(NN1)	2(NN2)	3(II)
pGR_{lex}	-1(PPIS2-we)	1(NN1)	2(NN2)	3(II-in)

Table 3: Pseudo-GR features for example sentence with window=3

Whichever feature set is used, an instance is simply the count of each GR's occurrences. We extract instances for the 385 verbs in the union of our two gold standards from the VALEX lexicon's data set, which was used in previous studies (Sun and Korhonen, 2009; Preiss et al., 2007) and facilitates comparison with that resource. This data set is drawn from five general-language corpora parsed by Rasp, and provides, on average, 7,000 instances per verb.

3.2 SCF extraction

Our graphical modeling approach uses the Bayesian network shown in Figure 1. Its generative story is as follows: when a verb is instantiated, an SCF is chosen according to a verb-specific multinomial. Then, the number and type of syntactic arguments (GRs) are chosen from two SCF-specific multinomials. These three multinomials are modeled with uniform Dirichlet priors and corresponding hyper-parameters α, β and γ. The model is trained via collapsed Gibbs sampling, where the probability of assigning a particular SCF s to an instance of verb v with GRs $(gr_1 \dots gr_n)$ is the product

$$\begin{aligned} P(s|Verb = v, GRs = gr_1 \dots gr_n) = \\ P(SCF = s|Verb = v) \times \\ P(N = n|SCF = s) \times \\ \prod_{i=1:n} P(GR = gr_i|SCF = s) \end{aligned}$$

The three terms, given the hyper-parameters and conjugate-prior relationship between Dirichlet and Multinomial distributions, can be expressed in terms of current assignments of s to verb v (c_{sv}), s to GR-count n (c_{sn}) and s to GR (c_{sg}), the corresponding totals (c_v, c_s), the dimensionality of the distributions ($|SCF|$, $|N|$ and $|G|$) and the hyper-parameters α, β and γ:

$$P(SCF = s|Verb = v) = (c_{sv}+\alpha)/(c_v+|SCF|\alpha)$$

$$P(N = n|SCF = s) = (c_{sn} + \beta)/(c_s + |N|\beta)$$

$$P(GR = gr_i|SCF = s) = (c_{sgr_i} + \gamma)/(c_s + |G|\gamma)$$

Note that N, the possible GR-count for an instance, is usually constant for pGRs ($2 \times window$), unless the verb is close to the start or end of the sentence.

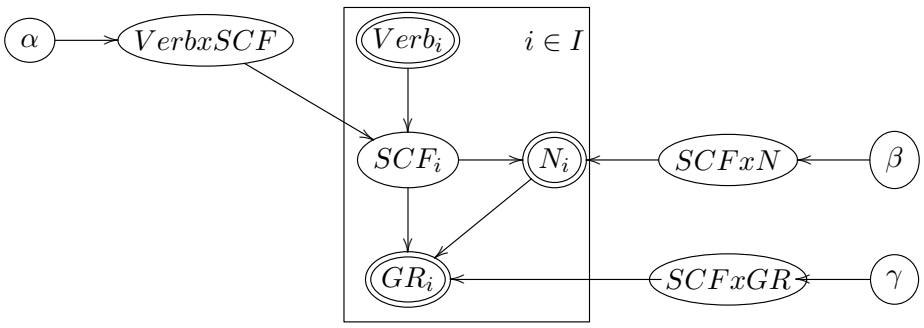

Figure 1: Our simple graphical model reflecting subcategorization. Double-circles indicate an observed value, arrows indicate conditional dependency. What constitutes a "GR" depends on the feature set being used.

We chose our hyper-parameters $\alpha = \beta = \gamma = .02$ to reflect the characteristic sparseness of the phenomena (i.e. verbs tend to take a small number of SCFs, which in turn are limited to a small number of realizations). For the pGRs we used a window of 5 tokens: a verb's arguments will fall within a small window in the majority of cases, so there is diminished return in expanding the window at the cost of increased noise. Finally, we set our SCF count to 40, about twice the size of the strictly syntactic general-language gold standard we describe in section 3.3. This overestimation allows some flexibility for the model to define its inventory based on the data; any supernumerary frames will act as "junk frames" that are rarely assigned and hence will have little influence. We run Gibbs sampling for 1000 iterations, and average the final 100 samples to estimate the posteriors $P(SCF|Verb)$ and $P(GR|SCF)$. Variance between adjacent states' estimates of $P(SCF|Verb)$ indicates that the sampling typically converges after about 100-200 iterations.[1]

3.3 Evaluation

Quantitative: cluster gold standard

Evaluating the output of unsupervised methods is not straightforward: discrete, expert-defined categories (like many SCF inventories) are unlikely to line up perfectly with data-driven, probabilistic output. Even if they do, finding a mapping between them is a problem of its own (Meila, 2003).

[1]Full source code for this work is available at `http://cl.cam.ac.uk/~tl318/files/subcat.tgz`

Our goal is to define a fair quantitative comparison between arbitrary SCF lexicons. An SCF lexicon makes two claims: first, that it defines a reasonable SCF inventory. Second, that for each verb, it has an accurate distribution over that inventory. We therefore compare the lexicons based on their performance on a task that a good SCF lexicon should be useful for: clustering verbs into lexical-semantic classes. Our gold standard is from (Sun and Korhonen, 2009), where 200 verbs were assigned to 17 classes based on their alternation patterns (Levin, 1993). Previous work (Schulte im Walde, 2009; Sun and Korhonen, 2009) has demonstrated that the quality of an SCF lexicon's inventory and probability estimates corresponds to its predictive power for membership in such alternation classes.

To compare the performance of our feature sets, we chose the simple and familiar K-Means clustering algorithm (Hartigan and Wong, 1979). The instances are the verbs' SCF distributions, and we select the number of clusters by the Silhouette validation technique (Rousseeuw, 1987). The clusters are then compared to the gold standard clusters with the purity-based F-Score from Sun and Korhonen (2009) and the more familiar Adjusted Rand Index (Hubert and Arabie, 1985). Our main point of comparison is the VALEX lexicon of SCF distributions, whose scores we report alongside ours.

Qualitative: manual gold standard

We also want to see how our results line up with a traditional linguistic view of subcategorization, but this requires digging into the unsupervised out-

put and associating anonymous probabilistic objects with established categories. We therefore present sample output in three ways: first, we show the clustering output from our top-performing method. Second, we plot the probability mass over GRs for two anonymous SCFs that correspond to recognizable traditional SCFs, and one that demonstrates unexpected behavior. Third, we compared the output for several verbs to a coarsened version of the manually-annotated gold standard used to evaluate VALEX (Preiss et al., 2007). We collapsed the original inventory of 168 SCFs to 18 purely syntactic SCFs based on their characteristic GRs and removed frames that depend on semantic distinctions, leaving the detection of finer-grained and semantically-based frames for future work.

4 Results

4.1 Verb clustering

We evaluated SCF lexicons based on the eight feature sets described in section 3.1, as well as the VALEX SCF lexicon described in section 2. Table 4 shows the performance of the lexicons in ascending order.

Method	Pur. F-score	Adj. Rand
tGR	.24	.02
tGR_{lim}	.27	.02
pGR_{lex}	.32	.09
$tGR_{lim,param}$	.35	.08
pGR	.35	.10
VALEX	.36	.10
$tGR_{param,lex}$	.37	.10
tGR_{param}	.39	.12
$tGR_{lim,param,lex}$	.44	.12

Table 4: Task-based evaluation of lexicons acquired with each of the eight feature types, and the state-of-the-art rule-based VALEX lexicon.

These results lead to several conclusions: first, training our model on tGRs outperforms pGRs and VALEX. Since the parser that produced them is known to perform well on general language (Briscoe et al., 2006), the tGRs are of high quality: it makes sense that reverting to the pGRs is unnecessary in this case. The interesting point is the major performance gain over VALEX, which uses the same tGR features along with expert-developed rules and inventory.

Second, we achieve performance comparable to VALEX using pGRs with a narrow window width. Since POS tagging is more reliable and robust across domains than parsing, retraining on new domains will not suffer the effects of a mismatched parsing model (Lippincott et al., 2010). It is therefore possible to use this method to build large-scale lexicons for any new domain with sufficient data.

Third, lexicalizing the closed-class POS tags introduces semantic information outside the scope of the alternation-based definition of subcategorization. For example, subdividing the indefinite pronoun tag "PN1" into "PN1-anyone" and "PN1-anything" gives information about the animacy of the verb's arguments. Our results show this degrades performance for both pGR and tGR features, unless the latter are limited to tGRs traditionally thought to be relevant for the task.

4.2 Qualitative analysis

Table 5 shows clusters produced by our top-scoring method, $GR_{param,lex,lim}$. Some clusters are immediately intelligible at the semantic level and correspond closely to the lexical-semantic classes found in Levin (1993). For example, clusters 1, 6, and 14 include member verbs of Levin's SAY, PEER and AMUSE classes, respectively. Some clusters are based on broader semantic distinctions (e.g. cluster 2 which groups together verbs related to locations) while others relate semantic classes purely based on their syntactic similarity (e.g. the verbs in cluster 17 share strong preference for 'to' preposition). The syntactic-semantic nature of the clusters reflects the multimodal nature of verbs and illustrates why a comprehensive subcategorization lexicon should not be limited to syntactic frames. This phenomenon is also encouraging for future work to tease apart and simultaneously exploit several verbal aspects via additional latent structure in the model.

An SCF's distribution over features can reveal its place in the traditional definition of subcategorization. Figure 2 shows the high-probability ($>$.02) tGRs for one SCF: the large mass centered on direct object tGRs indicates this approximates the notion of "transitive". Looking at the verbs most likely to take this SCF ("stimulate", "conserve") confirms

1	exclaim, murmur, mutter, reply, retort, say, sigh, whisper
2	bang, knock, snoop, swim, teeter
3	flicker, multiply, overlap, shine
4	batter, charter, compromise, overwhelm, regard, sway, treat
5	abolish, broaden, conserve, deepen, eradicate, remove, sharpen, shorten, stimulate, strengthen, unify
6	gaze, glance, look, peer, sneer, squint, stare
7	coincide, commiserate, concur, flirt, interact
8	grin, smile, wiggle
9	confuse, diagnose, march
10	mate, melt, swirl
11	frown, jog, stutter
12	chuckle, mumble, shout
13	announce, envisage, mention, report, state
14	frighten, intimidate, scare, shock, upset
15	bash, falter, snarl, wail, weaken
16	cooperate, eject, respond, transmit
17	affiliate, compare, contrast, correlate, forward, mail, ship

Table 5: Clusters (of size >2 and <20) produced using $tGR_{param,lex,lim}$

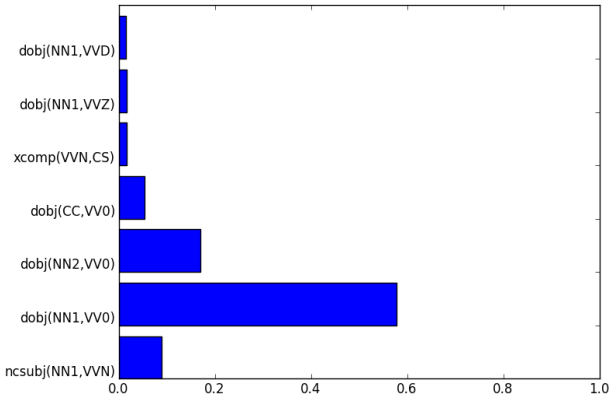

Figure 2: The SCF corresponding to transitive has most probability centered on dobj (e.g. stimulate, conserve, deepen, eradicate, broaden)

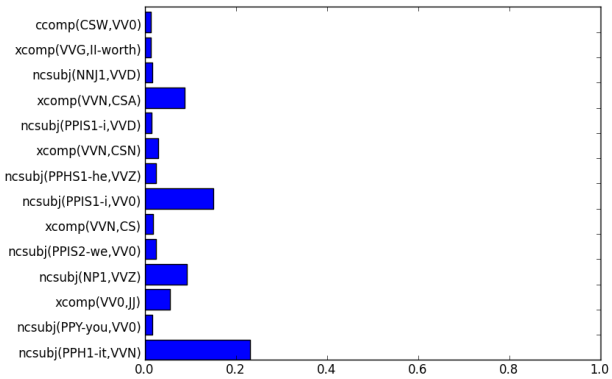

Figure 3: The SCF corresponding to verbs taking complements has more probability on xcomp and ccomp (e.g. believe, state, agree, understand, mention)

this. Figure 3 shows a complement-taking SCF, which is far rarer than simple transitive but also clearly induced by our model.

The induced SCF inventory also has some redundancy, such as additional transitive frames beside figure 2, and frames with poor probability estimates. Most of these issues can be traced to our simplifying assumption that each tGR is drawn independently w.r.t. an instance's other tGRs. For example, if an SCF gives *any* weight to indirect objects, it gives non-zero probability to an instance with *only* indirect objects, an impossible case. This can lead to skewed probability estimates: since some tGRs can occur multiple times in a given instance (e.g. indirect objects and prepositional phrases) the model may find it reasonable to create an SCF with all probability focused on that tGR, ignoring all others, such as in figure 4. We conclude that our independence assumption was too strong, and the model would benefit from defining more structure within instances.

The full tables necessary to compare verb SCF distributions from our output with the manual gold standard are prohibited by space, but a few examples reinforce the analysis above. The verbs "load" and "fill" show particularly high usage of ditransitive SCFs in the gold standard. In our inventory, this is reflected in high usage of an SCF with probability centered on indirect objects, but due to the independence assumptions the frame has a corresponding low probability on subjects and direct objects, despite the fact that these necessarily occur *along with* any indirect object. The verbs "acquire" and "buy" demonstrate both a strength of our approach and a weakness of using parsed input: both verbs

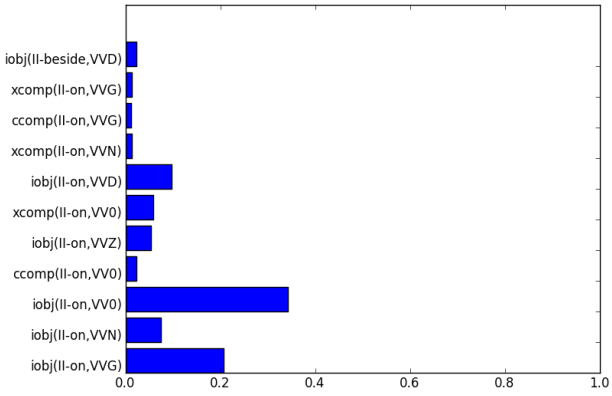

Figure 4: This SCF is dominated by indirect objects and complements, catering to verbs that may take several such tGRs, at the expense of subjects

show high probability of simple transitive in our output and the gold standard. However, the Rasp parser often conflates indirect objects and prepositional phrases due to its unlexicalized model. While our system correctly gives high probability to ditransitive for both verbs, it inherits this confusion and over-estimates "acquire"'s probability mass for the frame. This is an example of how bad decisions made by the parser cannot be fixed by the graphical model, and an area where pGR features have an advantage.

5 Conclusions and future work

Our study reached two important conclusions: first, given the same data as input, an unsupervised probabilistic model can outperform a hand-crafted rule-based SCF extractor with a predefined inventory. We achieve better results with far less effort than previous approaches by allowing the data to govern the definition of frames while estimating the verb-specific distributions in a fully Bayesian manner. Second, simply treating POS tags within a small window of the verb as pseudo-GRs produces state-of-the-art results without the need for a parsing model. This is particularly encouraging when building resources for new domains, where complex models fail to generalize. In fact, by integrating results from unsupervised POS tagging (Teichert and Daumé III, 2009) we could render this approach fully domain- and language-independent.

We did not dwell on issues related to choosing our hyper-parameters or latent class count. Both of these can be accomplished with additional sampling methods: hyper-parameters of Dirichlet priors can be estimated via slice sampling (Heinrich, 2009), and their dimensionality via Dirichlet Process priors (Heinrich, 2011). This could help address the redundancy we find in the induced SCF inventory, with the potential SCFs growing to accommodate the data.

Our initial attempt at applying graphical models to subcategorization also suggested several ways to extend and improve the method. First, the independence assumptions between GRs in a given instance turned out to be too strong. To address this, we could give instances internal structure to capture conditional probability between generated GRs. Second, our results showed the conflation of several verbal aspects, most notably the syntactic and semantic. In a sense this is encouraging, as it motivates our most exciting future work: augmenting this simple model to explicitly capture complementary information such as distributional semantics (Blei et al., 2003), diathesis alternations (McCarthy, 2000) and selectional preferences (Ó Séaghdha, 2010). This study targeted high-frequency verbs, but the use of syntactic and semantic classes would also help with data sparsity down the road. These extensions would also call for a more comprehensive evaluation, averaging over several tasks, such as clustering by semantics, syntax, alternations and selectional preferences.

In concrete terms, we plan to introduce latent variables corresponding to syntactic, semantic and alternation classes, that will determine a verb's syntactic arguments, their semantic realization (i.e. selectional preferences), and possible predicate-argument structures. By combining the syntactic classes with unsupervised POS tagging (Teichert and Daumé III, 2009) and the selectional preferences with distributional semantics (Ó Séaghdha, 2010), we hope to produce more accurate results on these complementary tasks while avoiding the use of any supervised learning. Finally, a fundamental advantage of a data-driven, parse-free method is that it can be easily trained for new domains. We next plan to test our method on a new domain, such as biomedical text, where verbs are known to take on distinct syntactic behavior (Lippincott et al., 2010).

6 Acknowledgements

The work in this paper was funded by the Royal Society, (UK), EPSRC (UK) grant EP/G051070/1 and EU grant 7FP-ITC-248064. We are grateful to Lin Sun and Laura Rimell for the use of their clustering and subcategorization gold standards, and the ACL reviewers for their helpful comments and suggestions.

References

Omri Abend and Ari Rappoport. 2010. Fully unsupervised core-adjunct argument classification. In *ACL '10*.

Galen Andrew, Trond Grenager, and Christopher Manning. 2004. Verb sense and subcategorization: using joint inference to improve performance on complementary tasks. *EMNLP '04*.

Collin Baker, Charles Fillmore, and John Lowe. 1998. The Berkeley FrameNet project. In *COLING ACL '98*.

David Blei, Andrew Ng, Michael Jordan, and John Lafferty. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32.

Bran Boguraev and Ted Briscoe. 1987. Large lexicons for natural language processing. *Computational Linguistics*, 13.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL on Interactive presentation sessions*.

John Carroll, Guido Minnen, and Ted Briscoe. 1998. Can subcategorisation probabilities help a statistical parser? In *The 6th ACL/SIGDAT Workshop on Very Large Corpora*.

K Bretonnel Cohen and Lawrence Hunter. 2006. A critical review of PASBio's argument structures for biomedical verbs. *BMC Bioinformatics*, 7.

James Curran, Stephen Clark, and Johan Bos. 2007. Linguistically motivated large-Scale NLP with C&C and Boxer. In *ACL '07*.

Marie-Catherine De Marneffe, Bill Maccartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *LREC '06*.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2007. The infinite tree. In *ACL '07*.

Ralph Grishman, Catherine Macleod, and Adam Meyers. 1994. Comlex syntax: building a computational lexicon. In *COLING '94*.

Xiwu Han, Chengguo Lv, and Tiejun Zhao. 2008. Weakly supervised SVM for Chinese-English cross-lingual subcategorization lexicon acquisition. In *The 11th Joint Conference on Information Science*.

J.A. Hartigan and M.A. Wong. 1979. Algorithm AS 136: A K-Means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*.

Gregor Heinrich. 2009. Parameter estimation for text analysis. Technical report, Fraunhofer IGD.

Gregor Heinrich. 2011. Infinite LDA implementing the HDP with minimum code complexity. Technical report, arbylon.net.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2.

Eric Joanis and Suzanne Stevenson. 2003. A general feature space for automatic verb classification. In *EACL '03*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. In *LREC '08*.

Anna Korhonen, Genevieve Gorrell, and Diana McCarthy. 2000. Statistical filtering and subcategorization frame acquisition. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.

Anna Korhonen, Yuval Krymolowski, and Ted Briscoe. 2006a. A large subcategorization lexicon for natural language processing applications. In *LREC '06*.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 2006b. Automatic classification of verbs in biomedical texts. In *ACL '06*.

Anna Korhonen, Yuval Krymolowski, and Nigel Collier. 2008. The choice of features for classification of verbs in biomedical texts. In *COLING '08*.

Ro Lenci, Barbara Mcgillivray, Simonetta Montemagni, and Vito Pirrelli. 2008. Unsupervised acquisition of verb subcategorization frames from shallow-parsed corpora. In *LREC '08*.

Beth Levin. 1993. *English Verb Classes and Alternation: A Preliminary Investigation*. University of Chicago Press, Chicago, IL.

Thomas Lippincott, Anna Korhonen, and Diarmuid Ó Séaghdha. 2010. Exploring subdomain variation in biomedical language. *BMC Bioinformatics*.

Diana McCarthy. 2000. Using semantic preferences to identify verbal participation in role switching alternations. In *NAACL '00*.

Marina Meila. 2003. Comparing clusterings by the Variation of Information. In *COLT*.

Paola Merlo and Suzanne Stevenson. 2001. Automatic verb classification based on statistical distributions of argument structure. *Computational Linguistics*.

Cédric Messiant. 2008. A subcategorization acquisition system for French verbs. In *ACL HLT '08 Student Research Workshop*.

Yusuke Miyao. 2005. Probabilistic disambiguation models for wide-coverage HPSG parsing. In *ACL '05*.

Radford M. Neal. 1993. Probabilistic inference using markov chain Monte Carlo methods. Technical report, University of Toronto.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan Mcdonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *The CoNLL Shared Task Session of EMNLP-CoNLL 2007*.

Diarmuid Ó Séaghdha. 2010. Latent variable models of selectional preference. In *ACL '10*.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics*.

Judita Preiss, Ted Briscoe, and Anna Korhonen. 2007. A system for large-scale acquisition of verbal, nominal and adjectival subcategorization frames from corpora. In *ACL '07*.

Douglas Roland and Daniel Jurafsky. 1998. How verb subcategorization frequencies are affected by corpus choice. In *ACL '98*.

Peter Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*.

C.J. Rupp, Paul Thompson, William Black, and John McNaught. 2010. A specialised verb lexicon as the basis of fact extraction in the biomedical domain. In *Interdisciplinary Workshop on Verbs: The Identification and Representation of Verb Features*.

Sabine Schulte im Walde. 2009. The induction of verb frames and verb classes from corpora. In *Corpus Linguistics. An International Handbook*. Mouton de Gruyter.

Lin Sun and Anna Korhonen. 2009. Improving verb clustering with automatically acquired selectional preferences. In *EMNLP'09*.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *ACL '03*.

Adam R. Teichert and Hal Daumé III. 2009. Unsupervised part of speech tagging without a lexicon. In *NIPS Workshop on Grammar Induction, Representation of Language and Language Learning*.

E. Uzun, Y. Klaslan, H.V. Agun, and E. Uar. 2008. Web-based acquisition of subcategorization frames for Turkish. In *The Eighth International Conference on Artificial Intelligence and Soft Computing*.

Giulia Venturi, Simonetta Montemagni, Simone Marchi, Yutaka Sasaki, Paul Thompson, John McNaught, and Sophia Ananiadou. 2009. Bootstrapping a verb lexicon for biomedical information extraction. In *Computational Linguistics and Intelligent Text Processing*. Springer Berlin / Heidelberg.