Semantic Roles for Nominal Predicates: Building a Lexical Resource

Ashwini Vaidya and **Martha Palmer** and **Bhuvana Narasimhan**
Dept of Linguistics
Institute of Cognitive Science
University of Colorado, Boulder
Boulder, CO 80309
{vaidyaa, mpalmer, narasimb}@colorado.edu

Abstract

The linguistic annotation of noun-verb complex predicates (also termed as light verb constructions) is challenging as these predicates are highly productive in Hindi. For semantic role labelling, each argument of the noun-verb complex predicate must be given a role label. For complex predicates, frame files need to be created specifying the role labels for each noun-verb complex predicate. The creation of frame files is usually done manually, but we propose an automatic method to expedite this process. We use two resources for this method: Hindi PropBank frame files for simple verbs and the annotated Hindi Treebank. Our method perfectly predicts 65% of the roles in 3015 unique noun-verb combinations, with an additional 22% partial predictions, giving us 87% useful predictions to build our annotation resource.

1 Introduction

Ahmed et al. (2012) describe several types of complex predicates that are found in Hindi e.g. morphological causatives, verb-verb complex predicates and noun-verb complex predicates. Of the three types, we will focus on the noun-verb complex predicates in this paper. Typically, a noun-verb complex predicate *chorii* 'theft' *karnaa* 'to do' has two components: a noun *chorii* and a light verb *karnaa* giving us the meaning 'steal'. Complex predicates [1] may be found in English e.g. *take a walk* and many other languages such as Japanese, Persian, Arabic and Chinese (Butt, 1993; Fazly and Stevenson, 2007).

[1]They are also otherwise known as light verb, support verb or conjunct verb constructions.

The verbal component in noun-verb complex predicates (NVC) has reduced predicating power (although it is inflected for person, number, and gender agreement as well as tense-aspect and mood) and its nominal complement is considered the true predicate, hence the term 'light verb'. The creation of a lexical resource for the set of true predicates that occur in an NVC is important from the point of view of linguistic annotation. For semantic role labelling in particular, similar lexical resources have been created for complex predicates in English, Arabic and Chinese (Hwang et al., 2010).

1.1 Background

The goal of this paper is to produce a lexical resource for Hindi NVCs. This resource is in the form of 'frame files', which are directly utilized for PropBank annotation. PropBank is an annotated corpus of semantic roles that has been developed for English, Arabic and Chinese (Palmer et al., 2005; Palmer et al., 2008; Xue and Palmer, 2003). In Hindi, the task of PropBank annotation is part of a larger effort to create a multi-layered treebank for Hindi as well as Urdu (Palmer et al., 2009).

PropBank annotation assumes that syntactic parses are already available for a given corpus. Therefore, Hindi PropBanking is carried out on top of the syntactically annotated Hindi Dependency Treebank. As the name suggests, the syntactic representation is dependency based, which has several advantages for the PropBank annotation process (see Section 3).

The PropBank annotation process for Hindi follows the same two-step process used for other PropBanks. First, the semantic roles that will occur with each predicate are defined by a human expert. Then,

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 126–131,
Atlanta, Georgia, 13-14 June 2013.

these definitions or 'frame files' are used to guide the annotation of predicate-argument structure in a given corpus.

Semantic roles are annotated in the form of *numbered arguments*. In Table 1 PropBank-style semantic roles are listed for the simple verb *de*;'to give':

de.01	'to give'
Arg0	the giver
Arg1	thing given
Arg2	recipient

Table 1: A frame file

The labels ARG0, ARG1 and ARG2 are always defined on a verb-by-verb basis. The description at the verb-specific level gives details about each numbered argument. In the example above, the numbered arguments correspond to the giver, thing given and recipient. In the Hindi treebank, which consists of 400,000 words, there are nearly 37,576 predicates, of which 37% have been identified as complex predicates at the dependency level. This implies that a sizeable portion of the predicates are NVCs, which makes the task of manual frame file creation time consuming.

In order to reduce the effort required for manual creation of NVC frame files, we propose a novel automatic method for generating PropBank semantic roles. The automatically generated semantic roles will be used to create frame files for each complex predicate in the corpus. Our method accurately predicts semantic roles for almost two thirds of the unique nominal-verb combinations, with around 20% partial predictions, giving us a total of 87% useful predictions.

For our implementation, we use linguistic resources in the form of syntactic dependency labels from the treebank. In addition we also have manually created, gold standard frame files for Hindi **simple** verbs[2]. In the following sections we provide linguistic background, followed by a detailed description of our method. We conclude with an error analysis and evaluation section.

[2]http://verbs.colorado.edu/propbank/framesets-hindi/

2 The Nominal and the Light Verb

Semantic roles for the arguments of the light verb are determined jointly by the noun as well as the light verb. Megerdoomian (2001) showed that the light verb places some restrictions on the semantic role of its subject in Persian. A similar phenomenon may be observed for Hindi. Compare example 1 with example 2 below:

(1) *Raam-ne cycle-kii chorii kii*
Ram-erg cycle-gen theft do.prf
'Ram stole a bicycle'

(2) *aaj cycle-kii chorii huii*
Today cycle-gen theft be.pres
'Today a bicycle was stolen'

PropBank annotation assumes that sentences in the corpus have already been parsed. The annotation task involves identification of arguments for a given NVC and the labelling of these arguments with semantic roles. In example 1 we get an agentive subject with the light verb *kar* 'do'. However, when it is replaced by the unaccusative *ho* 'become' in Example 2, then the resulting clause has a theme argument as its subject. Note that the nominal *chorii* in both examples remains the same. From the point of view of PropBank annotation, the NVC *chorii kii* will have both ARG0 and ARG1, but *chorii huii* will only have ARG1 for its single argument *cycle*. Hence, the frame file for a given nominal must make reference to the type of light verb that occurs with it.

The nominal as the true predicate also contributes its own arguments. In example 3, which shows a full (non-light) use of the verb *de* 'give', there are three arguments: giver(agent), thing given(theme) and recipient. In contrast the light verb usage *zor de* 'emphasis give; emphasize', seen in example 4, has a locative marked argument *baat par* 'matter on' contributed by the nominal *zor* 'emphasis'.

(3) *Raam-ne Mohan ko kitaab dii*
Ram-erg Mohan-dat book give.prf
'Ram gave Mohan a book'

(4) *Ram ne is baat par zor diyaa*
Ram-erg this matter loc emphasis give.prf
'Ram emphasized this matter'

As both noun and light verb contribute to the semantic roles of their arguments, we require linguistic knowledge about both parts of the NVC. The semantic roles for the nominal need to specify the co-occurring light verb and the nominal's argument roles must also be captured. Table 2 describes the desired representation for a nominal frame file.

Frame file for *chorii*-n(oun)	
chorii.01: theft-n	light verb: *kar*'do; to steal'
`Arg0` `Arg1`	person who steals thing stolen
chorii.02 : theft-n	light verb: *ho* 'be/become; to get stolen'
`Arg1`	thing stolen

Table 2: Frame file for predicate noun chorii 'theft' with two frequently occurring light verbs *ho* and *kar*. If other light verbs are found to occur, they are added as additional rolesets as chorii.03, chorii.04 and so on.

This frame file shows the representation of a nominal *chorii* 'theft' that can occur in combination with a light verb *kar* 'do' or *ho* 'happen'. For each combination, we derive a different set of PropBank roles: agent and patient for chorii.01 and theme for chorii.02. Note that the nominal's frame actually contains the roles for the combination of nominal and light verb, and not the nominal alone.

Nominal frame files such as these have already been defined for English PropBank.[3] However, for English, many nominals in NVCs are in fact nominalizations of full verbs, which makes it far easier to derive their frame files (e.g. *walk* in *take a walk* is a full verb). For Hindi, this is not the case, and a different strategy needs to be employed to derive these frames automatically.

3 Generating Semantic Roles

The Hindi Treebank has already identified NVC cases by using a special label `pof` or 'part-of'. The Treebank annotators apply this label on the basis of native speaker intuition. We use the label given by the Treebank as a means to extract the NVC cases (the issues related to complex predicate identification are beyond the scope of this paper). Once this extraction step is complete, we have a set of nominals and a corresponding list of light verbs that occur with them.

In Section 2, we showed that the noun as well as the light verb in a sentence influence the type of semantic roles that will occur. Our method builds on this idea and uses two resources in order to derive linguistic knowledge about the NVC: PropBank frame files for simple verbs in Hindi and the Hindi Treebank, annotated with dependency labels. The next two sections describe the use of these resources in some detail.

3.1 *Karaka* to PropBank Mapping

The annotated Hindi Treebank is based on a dependency framework (Begum et al., 2008) and has a very rich set of dependency labels. These labels (also known as *karaka* labels) represent the relations between a head (e.g. a verb) and its dependents (e.g. arguments). Using the Treebank we extract all the dependency *karaka* label combinations that occur with a unique instance of an NVC. We filter them to include argument labels and discard those labels that are usually used for adjuncts. We then calculate the most frequently occurring combination of labels that will occur with that NVC. Finally, we get a tuple consisting of an NVC, a set of *karaka* argument labels that occur with it and a count of the number of times that NVC has occurred in the corpus. The *karaka* labels are then mapped onto PropBank labels. We reproduce in Table 3 the numbered arguments to *karaka* label mapping found in Vaidya et al., (2011).

PropBank label	**Treebank label**
`Arg0` (agent)	`k1` (karta); `k4a` (experiencer)
`Arg1` (theme, patient)	`k2` (karma)
`Arg2` (beneficiary)	`k4` (beneficiary)
`Arg2-ATR`(attribute)	`k1s` (attribute)
`Arg2-SOU`(source)	`k5` (source)
`Arg2-GOL`(goal)	`k2p` (goal)
`Arg3` (instrument)	`k3` (instrument)

Table 3: Mapping from *Karaka* labels to PropBank

3.2 Verb Frames

Our second resource consists of PropBank frames for full Hindi verbs. Every light verb that occurs in

[3] http://verbs.colorado.edu/propbank/framesets-noun/

Hindi is also used as a full verb, e.g. *de* 'give' in Table 1 may be used both as a 'full' verb as well as a 'light' verb. As a full verb, it has a frame file in Hindi PropBank. The set of roles in the full verb frame is used to generate a "canonical" verb frame for each light verb. The argument structure of the light verb will change when combined with a nominal, which contributes its own arguments. However, as a default, the canonical argument structure list captures the fact that most *kar* 'do' light verbs are likely to occur with the roles `ARG0` and `ARG1` respectively or that *ho* 'become', an unaccusative verb, occurs with only `ARG1`.

3.3 Procedure

Our procedure integrates the two resources described above. First, the tuple consisting of *karaka* labels for a particular NVC is mapped to PropBank labels. But many NVC cases occur just once in the corpus and the *karaka* label tuple may not be very reliable. Hence, the likelihood that the mapped tuple accurately depicts the correct semantic frame is not very high. Secondly, Hindi can drop mandatory subjects or objects in a sentence e.g., *(vo) kitaab paRegaa*; '(He) will read the book'. These are not inserted by the dependency annotation (Bhatia et al., 2010) and are not easy to discover automatically (Vaidya et al., 2012). We cannot afford to ignore any of the low frequency cases as each NVC in the corpus must be annotated with semantic roles. In order to get reasonable predictions for each NVC, we use a simple rule. We carry out a mapping from karaka to PropBank labels only if the NVC occurs at least 30 times in the corpus. If the NVC occurs fewer than 30 times, then we use the "canonical" verb list.

4 Evaluation

The automatic method described in the previous section generated 1942 nominal frame files. In order to evaluate the frame files, we opted for manual checking of the automatically generated frames. The frame files were checked by three linguists and the checking focused on the validity of the semantic roles. The linguists also indicated whether annotation errors or duplicates were present. There was some risk that the automatically derived frames could bias the linguists' choice of roles as it is quicker to accept a given suggestion than propose an entirely new set of roles for the NVC. As we had a very large number of automatically generated frames, all of which would need to be checked manually anyway, practical concerns determined the choice of this evaluation.

After this process of checking, the total number of frame files stood at 1884. These frame files consisted of 3015 rolesets i.e. individual combinations of a nominal with a light verb (see Table 2). The original automatically generated rolesets were compared with their hand corrected counterparts (i.e. manually checked 'gold' rolesets) and evaluated for accuracy. We used three parameters to compare the gold rolesets with the automatically generated ones: a full match, partial match and no match. Table 4 shows the results derived from each resource (Section 3) and the total accuracy.

Type of Match	Full	Partial	None	Errors
Karaka Mapping	25	31	4	0
Verbal Frames	1929	642	249	143
Totals	1954	673	245	143
% Overall	65	22	8	5

Table 4: Automatic mapping results, total frames=3015

The results show that almost two thirds of the semantic roles are guessed correctly by the automatic method, with an additional 22% partial predictions, giving us a total of 87% useful predictions. Only 8% show no match at all between the automatically generated labels and the gold labels.

When we compare the contribution of the karaka labels with the verb frames, we find that the verb frames contribute to the majority of the full matches. The karaka mapping contributes relatively less as only 62 NVC types occur more than 30 times in the corpus. If we reduce our frequency requirement from of 30 to 5, the accuracy drops by 5%. The bulk of the cases are thus derived from the simple verb frames. We think that the detailed information in the verb frames, such as unaccusativity contributes towards generating the correct frame files.

It is interesting to observe that nearly 65% accuracy can be achieved from the verbal information alone. The treebank has two light verbs that occur with high frequency i.e. *kar* 'do' and *ho* 'become'. These combine with a variety of nominals but per-

Light verb	Full (%)	None (%)	Total Uses*
kar‘do’	**64**	8	1038
ho ‘be/become’	**81**	3	549
de ‘give’	55	**34**	157
A ‘come’	31	**42**	36

Table 5: Light verbs ‘do’ and ‘be/become’ vs. ‘give’ and ‘come’. *The unique total light verb usages in the corpus

form more consistently than light verbs such as *de* ‘give’ or *A* ‘come’. The light verb *kar* adds intentionality to the NVC, but appears less often with a set of semantic roles that are quite different from its original ‘full’ verb usage. In comparison, the light verbs such as *de* ‘give’ show far more variation, and as seen from Table 4, will match with automatically derived frames to a lesser extent. The set of nominals that occur in combination with *kar*, usually seem to require only a doer and a thing done. Borrowed English verbs such *dijain*‘design’ or *Pona*‘phone’ will appear preferentially with *kar* in the corpus and as they are foreign words they do not add arguments of their own.

One of the advantages of creating this lexical resource is the availability of gold standard frame files for around 3000 NVCs in Hindi. As a next step, it would be useful to use these frames to make some higher level generalizations about these NVCs. For example, much work has already been done on automatic verb classification for simple predicates e.g. (Merlo and Stevenson, 2001; Schulte im Walde, 2006), and perhaps such classes can be derived for NVCs. Also, the frame files do not currently address the problem of polysemous NVCs which could appear with a different set of semantic roles, which will be addressed in future work.

Acknowledgments

I am grateful to Archna Bhatia and Richa Srishti for their help with evaluating the accuracy of the nominal frames. This work is supported by NSF grants CNS-0751089, CNS-0751171, CNS-0751202, and CNS-0751213.

References

Tafseer Ahmed, Miriam Butt, Annette Hautli, and Sebastian Sulger. 2012. A reference dependency bank for analyzing complex predicates. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12*.

Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency Annotation Scheme for Indian Languages. In *Proceedings of The Third International Joint Conference on Natural Language Processing (IJCNLP). Hyderabad, India*.

Archna Bhatia, Rajesh Bhatt, Bhuvana Narasimhan, Martha Palmer, Owen Rambow, Dipti Misra Sharma, Michael Tepper, Ashwini Vaidya, and Fei Xia. 2010. Empty Categories in a Hindi Treebank. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC’10)*, pages 1863–1870.

Miriam Butt. 1993. The Light Verb Jungle. In G. Aygen, C. Bowers, and C. Quinn, editors, *Harvard Working Papers in Linguistics: Papers from the GSAS/Dudley House workshop on light verbs*, volume 9.

Afsaneh Fazly and Suzanne Stevenson. 2007. Automatic Acquisition of Knowledge about Multiword Predicates. In *Proceedings of PACLIC 19, the 19th Asia-Pacific Conference on Language, Information and Computation*.

Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. PropBank Annotation of Multilingual Light Verb Constructions. In *Proceedings of the Linguistic Annotation Workshop held in conjunction with ACL-2010*.

Karine Megerdoomian. 2001. Event Structure and Complex Predicates in Persian. *Canadian Journal of Linguistics*, 46:97–125.

Paola Merlo and Suzanne Stevenson. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–408.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Martha Palmer, Olga Babko-Malaya, Ann Bies, Mona Diab, Mohammed Maamouri, Aous Mansouri, and Wajdi Zaghouani. 2008. A pilot Arabic PropBank. In *Proceedings of the 6th International Language Resources and Evaluation*.

Martha Palmer, Rajesh Bhatt, Bhuvana Narasimhan, Owen Rambow, Dipti Misra Sharma, and Fei Xia. 2009. Hindi Syntax: Annotating Dependency, Lexical

Predicate-Argument Structure, and Phrase Structure. In *Proceedings of ICON-2009: 7th International Conference on Natural Language Processing*, Hyderabad.

Sabine Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

Ashwini Vaidya, Jinho D. Choi, Martha Palmer, and Bhuvana Narasimhan. 2011. Analysis of the Hindi proposition bank using dependency structure. In *Proceedings of the 5th Linguistic Annotation Workshop - LAW V '11*.

Ashwini Vaidya, Jinho D. Choi, Martha Palmer, and Bhuvana Narasimhan. 2012. Empty Argument Insertion in the Hindi PropBank. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation - LREC-12, Istanbul*.

Nianwen Xue and Martha Palmer. 2003. Annotating the Propositions in the Penn Chinese Treebank. In *Proceedings of the 2nd SIGHAN workshop on Chinese language processing*, SIGHAN'03, pages 47–54.