Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus

R. Mahesh K. Sinha
Department of Computer Science & Engineering
Indian Institute of Technology, Kanpur
Kanpur 208016 India
rmk@iitk.ac.in

Abstract

Complex predicate is a noun, a verb, an adjective or an adverb followed by a light verb that behaves as a single unit of verb. Complex predicates (CPs) are abundantly used in Hindi and other languages of Indo Aryan family. Detecting and interpreting CPs constitute an important and somewhat a difficult task. The linguistic and statistical methods have yielded limited success in mining this data. In this paper, we present a simple method for detecting CPs of all kinds using a Hindi-English parallel corpus. A CP is hypothesized by detecting absence of the conventional meaning of the light verb in the aligned English sentence. This simple strategy exploits the fact that CP is a multiword expression with a meaning that is distinct from the meaning of the light verb. Although there are several shortcomings in the methodology, this empirical method surprisingly yields mining of CPs with an average precision of 89% and a recall of 90%.

1 Introduction

Complex predicates (CPs) are abundantly used in Hindi and other languages of Indo-Aryan family and have been widely studied (Hook, 1974; Abbi, 1992; Verma, 1993; Mohanan, 1994; Singh, 1994; Butt, 1995; Butt and Geuder, 2001; Butt and Ramchand, 2001; Butt et al., 2003). A complex predicate is a multi-word expression (MWE) where a noun, a verb or an adjective is followed by a light verb (LV) and the MWE behaves as a single unit of verb. The general theory of complex predicate is discussed in Alsina (1996). These studies attempt to model the linguistic facts of complex predicate formation and the associated semantic roles.

CPs empower the language in its expressiveness but are hard to detect. Detection and interpretation of CPs are important for several tasks of natural language processing tasks such as machine translation, information retrieval, summarization etc. A mere listing of the CPs constitutes a valuable linguistic resource for lexicographers, wordnet designers (Chakrabarti et al., 2007) and other NLP system designers. Computational method using Hindi corpus has been used to mine CPs and categorize them based on statistical analysis (Sriram and Joshi, 2005) with limited success. Chakrabarti et al. (2008) present a method for automatic extraction of V+V CPs only from a corpus based on linguistic features. They report an accuracy of about 98%. An attempt has also been made to use a parallel corpus for detecting CPs using projection POS tags from English to Hindi (Soni, Mukerjee and Raina, 2006). It uses Giza++ word alignment tool to align the projected POS information. A success of 83% precision and 46% recall has been reported.

In this paper, we present a simple strategy for mining of CPs in Hindi using projection of meaning of light verb in a parallel corpus. In the following section the nature of CP in Hindi is outlined and this is followed by system design, experimentation and results.

2 Complex Predicates in Hindi

A CP in Hindi is a syntactic construction consisting of either a verb, a noun, an adjective or an adverb as main predicator followed by a light verb (LV). Thus, a CP can be a noun+LV, an adjective+LV, a verb+LV or an adverb+LV. Further, it is also possible that a CP is followed by a LV (CP+LV). The light verb carries the tense and agreement morphology. In V+V CPs, the contribution of the light verb denotes aspectual terms such as continuity, perfectivity, inception, completion, or denotes an expression of forcefulness, suddenness, etc. (Singh, 1994; Butt, 1995). The CP in a sentence syntactically acts as a single lexical unit of verb that has a meaning dis-

Proceedings of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009, pages 40–46,
Suntec, Singapore, 6 August 2009.

tinct from that of the LV. CPs are also referred as the complex or compound verbs.

Given below are some examples:

(1): CP=noun+LV
noun = *ashirwad* {blessings}
LV = *denaa* {to give}
usane mujhe ashirwad diyaa.
उसने मुझे आर्शीवाद दिया
{he me blessings gave}
he blessed me.

(2) No CP
usane mujhe ek pustak dii.
उसने मुझे एक पुस्तक दी
{he me one book gave}
he gave me a book.

In (1), the light verb *diyaa* (gave) in its past tense form with the noun *ashirwad* (blessings) makes a complex predicate verb form *ashirwad diyaa* (blessed) in the past tense form. The CP here is *ashirwad denaa* and its corresponding English translation is 'to bless'. On the other hand in example (2), the verb *dii* (gave) is a simple verb in past tense form and is not a light verb. Although, same Hindi verb *denaa* (to give) is used in both the examples, it is a light verb in (1) and a main verb in (2). Whether it acts as a light verb or not, depends upon the semantics of the preceding noun. However, it is observed that the English meaning in case of the complex predicate is not derived from the individual meanings of the constituent words. It is this observation that forms basis of our approach for mining of CPs.

(3) CP=adjective+LV
adjective=khush {happy}
LV=*karanaa* {to do}
usane mujhe khush kiyaa.
उसने मुझे खुश किया
{he me happy did}
he pleased me.

Here the Hindi verb *kiyaa* (did) is the past tense form of a light verb *karanaa* (to do) and the preceding word *khush* (happy) is an adjective. The CP here is *khush karanaa* (to please).

(4) CP=verb+LV
verb = *paRhnaa* {to read}
LV = *lenaa* {to take}
usane pustak paRh liyaa.
उसने पुस्तक पढ़ लिया
{he book read took}
he has read the book.

Here the Hindi verb *liyaa* (took) is the past tense form of the light verb *lenaa* (to take) and the preceding word *paRh* (read) is the verb *paRhnaa* (to read) in its stem form. The CP is *paRh lenaa* (to finish reading). In such cases the light verb acts as an aspectual /modal or as an intensifier.

(5) CP=verb+LV
verb = *phaadanaa* {to tear}
LV = *denaa* {to give}
usane pustak phaad diyaa.
उसने पुस्तक फाड़ दिया
{he book tear gave}
he has torn the book.

Here the Hindi verb *diyaa* (gave) is the past tense form of the light verb *denaa* (to give) and the preceding word *phaad* (tear) is the stem form of the verb *phaadanaa* (to tear) . The CP is *phaad denaa* (to cause and complete act of tearing).

(6) CP=verb+LV
verb = *denaa* {to give}
LV = *maaranaa*{to hit/ to kill}
usane pustak de maaraa.
उसने पुस्तक दे मारा
{he book give hit}
he threw the book.

Here the Hindi verb *maaraa* (hit/killed) is the past tense form of the light verb *maaranaa* (to hit/ to kill) and the preceding word *de* (give) is a verb *denaa* (to give) in its stem form. The CP is *de maranaa* (to throw). The verb combination yields a new meaning. This may also be considered as a semi-idiomatic construct by some people.

(7) CP=adverb+LV1+LV2
adverb = *vaapas*{back}
LV1 = *karanaa*{to do}
LV2 = *denaa*{to give}
or
CP = CP+LV
CP = *vaapas karanaa*{to return}
LV = *denaa*{to give}
usane pustak vaapas kar diyaa.
उसने पुस्तक वापस कर दिया
{he book back do gave}

he returned the book.

Here there are two Hindi light verbs used. The verb *kar* (do) is the stem form of the light verb *karanaa* (to do) and the verb *diyaa* (gave) is the past tense form of the light verb *denaa* (to give). The preceding word *vaapas* (back) is an adverb. One way of interpretation is that the CP (a conjunct verb) *vaapas karanaa* (to return) is followed by another LV *denaa* (to give) signifying completion of the task. Another way of looking at it is to consider these as a combination of two CPs, *vaapas karanaa* (to return) and *kar denaa* (to complete the act). The semantic interpretations in the two cases remain the same. It may be noted that the word *vaapas* (return) is also a noun and in such a case the CP is a noun+LV.

From all the above examples, the complexity of the task of mining the CPs is evident. However, it is also observed that in the translated text, the meaning of the light verb does not appear in case of CPs. Our methodology for mining CPs is based on this observation and is outlined in the following section.

3 System Design

As outlined earlier, our method for detecting a CP is based on detecting a mismatch of the Hindi light verb meaning in the aligned English sentence. The steps involved are as follows:

1) Align the sentences of Hindi-English corpus;
2) Create a list of Hindi light verbs and their common English meanings as a simple verb; (Table 1)
3) For each Hindi light verb, generate all the morphological forms (Figure 1);
4) For each English meaning of the light verb as given in table 1, generate all the morphological forms (Figure 2);
5) For each Hindi-English aligned sentence, execute the following steps:
   a) For each light verb of Hindi (table 1), execute the following steps:
      i) Search for a Hindi light verb (LV) and its morphological derivatives (figure 1) in the Hindi sentence and mark its position in the sentence (K);
      ii) If the LV or its morphological derivative is found, then search for the equivalent English meanings for any of the morphological forms (figure 2) in the corresponding aligned English sentence;
      iii) If no match is found, then scan the words in the Hindi sentence to the left of the K^{th} position (as identified in step (i)); else if a match is found, then exit {i.e. go to step (a)}.
      iv) If the scanned word is a 'stop word' (figure 3), then ignore it and continue scanning;
      v) Stop the scan when it is not a 'stop word' and collect the Hindi word (W);
      vi) If W is an 'exit word' then exit {i.e. go to step (a)}, else the identified CP is W+LV.

Hindi has a large number of light verbs. A list of some of the commonly used light verbs along with their common English meanings as a simple verb is given in table 1. The light verb *kar* (do) is the most frequently used light verb. Using its literal meaning as 'do', as a criterion for testing CP is quite misleading since 'do' in English is used in several other contexts. Such meanings have been shown within parentheses and are not used for matching.

light verb base form	root verb meaning
baithanaa बैठना	sit
bananaa बनना	make/become/build/construct/ manufacture/prepare
banaanaa बनाना	make/build/construct/manufacture/ prepare
denaa देना	give
lenaa लेना	take
paanaa पाना	obtain/get
uthanaa उठना	rise/ arise/ get-up
uthaanaa उठाना	raise/lift/ wake-up
laganaa लगना	feel/appear/ look /seem
lagaanaa लगाना	fix/install/ apply
cukanaa चुकना	(finish)
cukaanaa चुकाना	pay
karanaa करना	(do)
honaa होना	happen/become /be
aanaa आना	come
jaanaa जाना	go
khaanaa खाना	eat
rakhanaa रखना	keep / put
maaranaa मारना	kill/beat/hit
daalanaa डालना	put
haankanaa हाँकना	drive

Table 1. Some of the common light verbs in Hindi

For each of the Hindi light verb, all morphological forms are generated. A few illustrations are given in figures 1(a) and 1(b). Similarly, for each of the English meaning of the light verb, all of its morphological derivatives are generated. Figure 2 shows a few illustrations of the same.

There are a number of words that can appear in between the nominal and the light verb in a CP. These words are ignored in search for a CP and are treated as stop words. These are words that denote negation or are emphasizers, intensifiers, interrogative pronoun or a particle. A list of stop words used in the experimentation is given in figure 3.

LV: jaanaa जाना {to go}
Morphological derivatives:
jaa jaae jaao jaae.M jaauu.M jaane jaanaa jaanii jaataa jaatii jaate jaanii.M jaatii.M jaaoge jaaogii gaii jaauu.MgA jaayegaa jaauu.Mgii jaayegii gaye gaii.M gayaa gayii jaaye.Mge jaaye.MgI jaakara
जा (go: stem) जाए (go: imperative)
जाओ (go: imperative) जाएं (go: imperative)
जाऊँ (go: first-person) जाने (go: infinitive, oblique)
जाना (go: infinitive, masculine, singular)
जानी (go: infinitive, feminine, singular)
जाता (go: indefinite, masculine, singular)
जाती (go: indefinite, feminine, singular)
जाते (go: indefinite, masculine, plural/oblique)
जानीं (go: infinitive, feminine, plural)
जातीं (go: indefinite, feminine, plural)
जाओगे (go: future, masculine, singular)
जाओगी (go: future, feminine, singular)
गई (go: past, feminine, singular)
जाऊँगा (go: future, masculine, first-person, singular)
जायेगा (go: future, masculine, third-person, singular)
जाऊँगी (go: future, feminine, first-person, singular)
जायेगी (go: future, feminine, third-person, singular)
गये (go: past, masculine, plural/oblique)
गईं (go: past, feminine, plural)
गया (go: past, masculine, singular)
गयी (go: past, feminine, singular)
जायेंगे (go: future, masculine, plural)
जायेंगी (go: future, feminine, plural)
जाकर (go: completion)
ooooo

Figure 1(a). Morphological derivatives of sample Hindi light verb 'jaanaa' जाना {to go}

LV: lenaa लेना {to take}
Morphological derivatives:
le lii le.M lo letaa letii lete lii.M luu.M legaa legii lene lenaa lenii liyaa le.Mge loge letii.M luu.Mgaa luu.Mgii lekara
ले (take: stem) ली (take: past)
लें (take: imperative) लो (take: imperative)
लेता (take: indefinite, masculine, singular)
लेती (take: indefinite, feminine, singular)
लेते (take: indefinite, masculine, plural/oblique)
लीं(take:past,feminine,plural) लूँ(take: first-person)
लेगा(take: future, masculine, third-person,singular)
लेगी(take: future, feminine, third-person, singular)
लेने (take: infinitive, oblique)
लेना (take: infinitive, masculine, singular)
लेनी (take: infinitive, feminine, singular)
लिया (take: past, masculine, singular)
लेंगे (take: future, masculine, plural)
लोगे (take: future, masculine, singular)
लेतीं (take: indefinite, feminine, plural)
लूँगा (take: future, masculine,first-person,singular)
लूँगी (take: future, feminine, first-person, singular)
लेकर (take: completion)
ooooo

Figure 1(b). Morphological derivatives of sample Hindi light verb 'lenaa' लेना {to take}

English word: sit
Morphological derivations:
sit sits sat sitting
English word: give
Morphological derivations:
give gives gave given giving
ooooo

Figure 2. Morphological derivatives of sample English meanings

We use a list of words of words that we have named as 'exit words' which cannot form part of a CP in Hindi. We have used Hindi case (*vibhakti*) markers (also called *parsarg*), conjunctions and pronouns as the 'exit words' in our implementation. Figure 4 shows a partial list used. However, this list can be augmented based on analysis of errors in LV identification. It should be noted that we do not perform parts of speech (POS) tagging and so the nature of the word preceding the LV is unknown to the system.

नहीं (no/not),
न (no/not /Hindi particle),
भी (also /Hindi particle),
ही(only /Hindi particle),
तो (then /Hindi particle),
क्यों (why),
क्या (what /Hindi particle),
कहाँ (where /Hindi particle),
कब (when),
यहाँ (here),
वहां (there),
जहाँ (where),
पहले (before),
बाद में (after),
शुरू में (beginning),
आरम्भ में (beginning),
अंत में (in the end),
आखिरी में (in the end).

Figure 3. Stop words in Hindi used by the system

ने (ergative case marker), को (accusative case marker), का (possessive case marker), के (possessive case marker), की (possessive case marker), से (from/by/with), में (in/into), पर (on/but), और (and/ Hindi particle), तथा (and), या (or), लेकिन (but), परन्तु (but), कि (that/ Hindi particle), मैं (I), तुम (you), आप (you), वह (he/she), मेरा (my), मेरी (my), मेरे (my), तुम्हारा (your), तुम्हारी (your), तुम्हारे (your), उसका (his), उसकी (her), उसके (his/her), अपना (own), अपनी (own), अपने (own), उनके (their), मैंने (I ergative), तुम्हे (to you), आपको (to you), उसको (to him/her), उनको (to them), उन्हें (to them), मुझको (to me), मुझे (to me), जिसका (whose), जिसकी (whose), जिसके (whose), जिनको (to whom), जिनके (to whom)

Figure 4. A few exit words in Hindi used by the system

The inner loop of the procedure identifies multiple CPs that may be present in a sentence. The outer loop is for mining the CPs in the entire corpus. The experimentation and results are discussed in the following section.

4 Experimentation and Results

The CP mining methodology outlined earlier has been implemented and tested over multiple files of EMILLE (McEnery, Baker, Gaizauskas and Cunningham, 2000) English-Hindi parallel corpus. A summary of the results obtained are given in table 2. As can be seen from this table, the precision obtained is 80% to 92% and the recall is between 89% to 100%. The F-measure is 88% to 97%. This is a remarkable and somewhat surprising result from the simple methodology without much of linguistic or statistical analysis. This is much higher than what has been reported on the same corpus by Mukerjee et al, 2006 (83% precision and 46% recall) who use projection of POS and word alignment for CP identification. This is the only other work that uses a parallel corpus and covers all kinds of CPs. The results as reported by Chakrabarti et al. (2008) are only for V-V CPs. Moreover they do not report the recall value.

	File 1	File 2	File 3	File 4	File 5	File 6
No. of Sentences	112	193	102	43	133	107
Total no. of CP(N)	200	298	150	46	188	151
Correctly identified CP (TP)	195	296	149	46	175	135
V-V CP	56	63	9	6	15	20
Incorrectly identified CP (FP)	17	44	7	11	16	20
Unidentified CP (FN)	5	2	1	0	13	16
Accuracy %	97.50	99.33	99.33	100,0	93.08	89.40
Precision % (TP/ (TP+FP))	91.98	87.05	95.51	80.70	91.62	87.09
Recall % (TP / (TP+FN))	97.50	98.33	99.33	100.0	93.08	89.40
F-measure % (2PR / (P+R))	94.6	92.3	97.4	89.3	92.3	88.2

Table 2. Results of the experimentation

Given below are some sample outputs:

(1)
English sentence:
I also enjoy working with the children's parents who often come to me for advice - it's good to know you can help.

Aligned Hindi sentence:
मुझे बच्चों के माता - पिताओं के साथ काम करना भी अच्छा लगता है जो कि अक्सर सलाह लेने आते हैं - यह जानकार खुशी होती है कि आप किसी की मदद कर सकते हैं |

The CPs identified in the sentence:
i. काम करना (to work), ii. अच्छा लगना (to feel good: enjoy), iii. सलाह लेना (to seek advice), iv. खुशी होना (to feel happy: good), v. मदद करना (to help)

Here the system identified 5 different CPs all of which are correct and no CP in the sentence has gone undetected. The POS projection and word alignment method (Mukerjee et al., 2006) would fail to identify CPs सलाह लेना (to seek advice), and खुशी होना (to feel happy).

(2)
English sentence:
Thousands of children are already benefiting from the input of people like you - people who care about children and their future, who have the commitment, energy and enthusiasm to be positive role models, and who value the opportunity for a worthwhile career.

Aligned Hindi sentence:
आप जैसे लोग जो कि बच्चों और उनके भविष्य के बारे में सोचते हैं - इस समय भी हज़ारों बच्चों को लाभ पहुँचा रहे हैं | अच्छे आदर्श बनने के लिए ऐसे लोगों में प्रतिबद्धता , उत्सराह और लगन है और वे एक समर्थन - योग्य व्यवसाय की कद्र करते हैं |

The CPs identified in the sentence:
i. आदर्श बनना (to be role model), ii. कद्र करना (to respect)

Here also the two CPs identified are correct.

It is obvious that this empirical method of mining CPs will fail whenever the Hindi light verb maps on to its core meaning in English. It may also produce garbage as POS of the preceding word is not being checked. However, the mining success rate obtained speaks of these being in small numbers in practice. Use of the 'stop words' in allowing the intervening words within the CPs helps a lot in improving the performance. Similarly, use of the 'exit words' avoid a lot of incorrect identification.

5 Conclusions

The simple empirical method for mining CPs outlined in this work, yields an average 89% of precision and 90% recall which is better than the results reported so far in the literature. The major drawback is that we have to generate a list of all possible light verbs. This list appears to be very large for Hindi. Since no POS tagging or statistical analysis is performed, the identified CPs are merely a list of mined CPs in Hindi with no linguistic categorization or analysis. However, this list of mined CPs is valuable to the lexicographers and other language technology developers. This list can also be used for word alignment tools where the identified components of CPs are grouped together before the word alignment process. This will increase both the alignment accuracy and the speed.

The methodology presented in this work is equally applicable to all other languages within the Indo-Aryan family.

References

Anthony McEnery, Paul Baker, Rob Gaizauskas, Hamish Cunningham. 2000. EMILLE: Building a Corpus of South Asian Languages, *Vivek, A Quarterly in Artiificial Intelligence,* 13(3):23–32.

Amitabh Mukerjee, Ankit Soni, and Achala M. Raina, 2006. Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora**,** *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties,* Sydney, 11–18,

Alex Alsina. 1996. *Complex Predicates:Structure and Theory*. CSLI Publications,Stanford, CA.

Anvita Abbi. 1992. The explicator compound verb:some definitional issues and criteria for identification. *Indian Linguistics*, 53, 27-46.

Debasri Chakrabarti, Vaijayanthi Sarma and Pushpak Bhattacharyya. 2007. Complex Predicates in Indian Language Wordnets, *Lexical Resources and Evaluation Journal*, 40 (3-4).

Debasri Chakrabarti, Hemang Mandalia, Ritwik Priya, Vaijayanthi Sarma and Pushpak Bhattacharyya. 2008. Hindi Compound Verbs and their Automatic Extraction, *Computational Linguistics (COLING08),* Manchester, UK.

Manindra K. Verma (ed.) 1993. *Complex Predicates in South Asian Languages.* Manohar Publishers and Distributors, New Delhi

Miriam Butt. 1995. *The Structure of Complex Predicates in Urdu.* CSLI Publications.

Mirium Butt and Gillian Ramchand. 2001. *Complex Aspectual Structure in Hindi/Urdu*. In Maria Liakata, Britta Jensen and Didier Maillat (Editors), Oxford University Working Papers in Linguistics, Philology & Phonetics, Vol. 6.

Miriam Butt, Tracy Holloway King, and John T. Maxwell III. 2003. Complex Predicates via Restriction, *Proceedings of the LFG03 Conference*.

Miriam Butt and Wilhelm Geuder. 2001. *On the (semi)lexical status of light verbs*. In Norbert Corver and Henk van Riemsdijk, (Editors), Semi-lexical Categories: On the content of function words and the function of content words, Mouton de Gruyter, Berlin, 323–370.

Mona Singh. 1994. *Perfectivity, Definiteness, and Specificity: A Classification of Verbal Predicates Hindi*. Doctoral dissertation, University of Texas, Austin.

Peter Edwin Hook. 1974. *The Compound Verb in Hindi*. Center for South and Southeast Asian Studies: The University of Michigan.

Tara Mohanan. 1994. *Argument Structure in Hindi*. CSLI Publications, Stanford, California

Venkatapathy Sriram and Aravind K. Joshi, 2005. Relative compositionality of multi-word expressions: a study of verb-noun (V-N) collocations, *In Proceedings of International Joint Conference on Natural Language Processing - 2005, Jeju Island, Korea*, 553-564.