Detecting Complex Predicates in Hindi across Parallel Corpora

CS671: Introduction to Natural Language Processing
Guide: Prof. Amitabha Mukerjee

Amit Kumar (10074) and Ankit Modi (10104)

Abstract

A Complex Predicate (CP) is a multi word compound that behaves as a single verb. One component of the compound is a light verb, which indicates tense, mood, or aspect, but provides only fine shades of meaning. The other, "primary", component is a verb or noun which carries most of the semantics of the compound, and determines its arguments. A Light Verb is a verb that has little semantic content of its own and it therefore forms a predicate with some additional expression, which is usually a noun.

This project aims at detecting CPs from a parallel English-Hindi corpora using the fact that a CP is a multi-word expression with its meaning being distinct from the light verb (LV).

Motivation

Complex predicate play a pivotal role in improving the expressiveness of a language. The abundance of CPs in Hindi makes detecting them, an important task. Moreover, detection of CPs would provide an important resource for tasks such as Wordnet construction, Linguistic analysis etc.

Previous Work

This problem was solved using word alignment and POS tagging of parallel sentences [2]. In this approach POS tagging of English sentence is projected on the parallel hindi sentence. Then by using word alignment of both sentences CPs are identified. CPs had been mined using computational methods and then, were categorized using statistical analysis (Sriram and Joshi, 2005). Chakrabarti et al (2008) present a method for automatic extraction of CPs only from a corpus based on linguistic features.

Derivation of complex predicates has also been dealt with linguistically and
computationally [3]. Authors have constructed empirical tests to decide if a combination of two words, the second of which is a verb, is a complex predicate or not. Such tests will provide a principled way of deciding the status of complex predicates in Indian language wordnets.

**Introduction**

Given a parallel English-Hindi corpora, we have to detect complex predicates from it. We would be detecting them by absence of the conventional meaning of the light verb in the aligned English sentence.

For example, consider the aligned sentences below,

* I also enjoy working with the children's parents who often come to me for advice - it's good to know you can help

* मुझे के माता-पिताओं के साथ काम करना भी अच्छा लगता है जो कि अक्सर सलाह लेने आते हैं - यह जानकार ख़शी होती है कि आप किसी कि मदद कर सकते हैं |

Here, apart from सकते (can) and आते (come), none of the light verbs have their conventional meanings in the English sentence. Thus, are contenders for constituting a Complex Predicate.

This project is mainly based on the paper “Mining Complex Predicates in Hindi Using a Parallel Hindi-English Corpus” by R. Mahesh and K. Sinha [1]. We have also consulted the paper “Detecting Complex Predicates in Hindi using POS projection across parallel corpora” [2] to compare our results. [1] evaluated their results on multiple files (exact numbers are not mentioned) whereas [2] detected 1439 instances of CPs from 4400 sentences. We would be evaluating this project on 0.1 million sentences.

**Our Work**

As we had mentioned earlier, this method is based on detecting a mismatch of the Hindi light verb meaning in the English aligned sentence. Following are the basic steps involved in the method:

- Align the sentences of English-Hindi corpus.
- Make a list of Hindi Light Verbs and their English meanings. See the figure below for example:

![Hindi Light Verbs and their English meanings](image1.png)

- For each Hindi Light Verb, generate all the morphological forms. For example:

![Morphological forms of Hindi Light Verbs](image2.png)

- For each English meaning as given in Fig1, generate all the morphological forms. For example:

![Morphological forms of English verbs](image3.png)
For each English-Hindi aligned sentence, do the following:

1. Search for a light verb or its morphological forms, say LV, in the Hindi sentence and mark its position, say K.

   For each of the light verbs found, do the following:

   i. Search for the Equivalent English meaning in the aligned English sentence.

   ii. If no match is found, then scan the words in the Hindi sentence to the left of the Kth position (as identified in step (1)); else if a match is found, then exit {i.e. go to step (i)}.

   iii. If the scanned word is a stop word, then ignore it and continue scanning.

   iv. Stop scanning if it is not a stop word and collect the Hindi word, say W.

   v. If W is an exit word, go to step (i), else W + LV is our CP

A flowchart for the algorithm would be:

![Flowchart of the algorithm](img/fig4.png)
Now, let us go through our algorithm with an example:

- Aligned English- Hindi corpus

  I also enjoy working with the children's parents who often come to me for advice. मुझे बच्चों के माता-पिताओं के साथ काम करना भी अच्छा लगता है जो कि अक्सर सलाह लेने आते हैं।

- Search for Hindi LV & its morphological forms

  I also enjoy working with the children's parents who often come to me for advice. मुझे बच्चों के माता-पिताओं के साथ काम करना भी अच्छा लगता है जो कि अक्सर सलाह लेने आते हैं।

- Search for equivalent English meaning of LVs

  I also enjoy working with the children's parents who often come to me for advice. मुझे बच्चों के माता-पिताओं के साथ काम करना भी अच्छा लगता है जो कि अक्सर सलाह लेने आते हैं।

- Scan left of those LVs whose English meaning is not found

  I also enjoy working with the children's parents who often come to me for advice. मुझे बच्चों के माता-पिताओं के साथ काम करना भी अच्छा लगता है जो कि अक्सर सलाह लेने आते हैं।

- Collect the Hindi word (W) if it is not a stop word or else keep scanning.

  I also enjoy working with the children's parents who often come to me for advice. मुझे बच्चों के माता-पिताओं के साथ काम करना भी अच्छा लगता है जो कि अक्सर सलाह लेने आते हैं।

- \( CP = W + LV, \) unless \( W \) is an exit word

  I also enjoy working with the children's parents who often come to me for advice. मुझे बच्चों के माता-पिताओं के साथ काम करना भी अच्छा लगता है जो कि अक्सर सलाह लेने आते हैं।

**Results**

The dataset we used had 61,000 aligned sentences. The mentioned method extracted 90,707 Complex Predicates from it. Out of these CPs, 24,893 were unique.

For Precision and Recall, we made 5 test files and manually calculated all the complex
predicates from it to compare them with the results obtained by this method. Following table summarises our result:

<table>
<thead>
<tr>
<th>File No.</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F_Score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>File 1</td>
<td>100</td>
<td>87.5</td>
<td>93.3</td>
</tr>
<tr>
<td>File 2</td>
<td>95.5</td>
<td>77.8</td>
<td>85.7</td>
</tr>
<tr>
<td>File 3</td>
<td>96.5</td>
<td>84</td>
<td>89.8</td>
</tr>
<tr>
<td>File 4</td>
<td>100</td>
<td>84</td>
<td>91.3</td>
</tr>
<tr>
<td>File 5</td>
<td>100</td>
<td>85</td>
<td>91.9</td>
</tr>
</tbody>
</table>

So we see that the precision lies in the range of 95-100 % whereas the range of recall is 77-88%. F_score lies in 85-93%.

Results in [1] were: Precision in 80-92% while Recall in 89-100%. F_score in 88-97%.
Results in [2] were: Precision of 82.5% while Recall in 40-46%. F_score in 53-59%.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
<th>F_score (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>95-100</td>
<td>77-88</td>
<td>85-93</td>
</tr>
<tr>
<td>[1]</td>
<td>80-92</td>
<td>89-100</td>
<td>88-97</td>
</tr>
<tr>
<td>[2]</td>
<td>82.5</td>
<td>40-46</td>
<td>53-59</td>
</tr>
</tbody>
</table>

We also calculated the frequency of each light verb (base forms). These frequencies ranged from 26,049 (कर) to 5 (टपक) 

Following is a bar graph showing the light verbs with the top 10 frequencies.
Future Work / Improvements:
The main drawback that we came across in this method is as follows:
Homophones: Words with same spelling as light verbs but different meanings were recognised as light verbs. For example:
मै बस मैं बैठा।
In this sentence, “मै बस” was detected as a CP as it has the same spelling as “बस” meaning inhabit”, which is a light verb.
Resources

- English- Hindi parallel Corpora (0.1 million sentences):
- List of Hindi Light Verbs: Reverse Complex Predicates in Hindi by Shakthi Poornima, Department of Linguistics, SUNY University at Buffalo
- Morphological forms of English verbs:
  http://www.englishpage.com/irregularverbs/irregularverbs.html
- Morphological forms of Hindi verbs:
  Extracted from the large Hindi corpus (Blog Corpus)

References

- Wikipedia: