Parsing Aligned Parallel Corpus by Projecting Syntactic Relations from Annotated Source Corpus

Shailly Goyal **Niladri Chatterjee**
Department of Mathematics
Indian Institute of Technology Delhi
Hauz Khas, New Delhi - 110 016, India
{shailly_goyal, niladri_iitd}@yahoo.com

Abstract

Example-based parsing has already been proposed in literature. In particular, attempts are being made to develop techniques for language pairs where the source and target languages are different, e.g. Direct Projection Algorithm (Hwa et al., 2005). This enables one to develop parsed corpus for target languages having fewer linguistic tools with the help of a resource-rich source language. The DPA algorithm works on the assumption of *Direct Correspondence* which simply means that the relation between two words of the source language sentence can be projected directly between the corresponding words of the parallel target language sentence. However, we find that this assumption does not hold good all the time. This leads to wrong parsed structure of the target language sentence. As a solution we propose an algorithm called pseudo DPA (pDPA) that can work even if Direct Correspondence assumption is not guaranteed. The proposed algorithm works in a recursive manner by considering the embedded phrase structures from outermost level to the innermost. The present work discusses the pDPA algorithm, and illustrates it with respect to English-Hindi language pair. Link Grammar based parsing has been considered as the underlying parsing scheme for this work.

1 Introduction

Example-based approaches for developing parsers have already been proposed in literature. These approaches either use examples from the same language, e.g., (Bod et al., 2003; Streiter, 2002), or they try to imitate the parse of a given sentence using the parse of the corresponding sentence in some other language (Hwa et al., 2005; Yarowsky and Ngai, 2001). In particular, Hwa et al. (2005) have proposed a scheme called *direct projection algorithm* (DPA) which assumes that the relation between two words in the source language sentence is preserved across the corresponding words in the parallel target language. This is called Direct Correspondence Assumption (DCA).

However, with respect to Indian languages we observed that the DCA does not hold good all the time. In order to overcome the difficulty, in this work, we propose an algorithm based on a variation of the DCA, which we call *pseudo Direct Correspondence Assumption* (pDCA). Through pDCA the syntactic knowledge can be transferred even if not all syntactic relations may be projected directly from the source language to the target language in toto. Further, the proposed algorithm projects the relations between phrases instead of projecting relations between words. Keeping in line with (Hwa et al., 2005), we call this algorithm as *pseudo Direct Projection Algorithm* (pDPA).

The present work discusses the proposed parsing scheme for a new (target) language with the help of a parser that is already available for a language (source) and using word-aligned parallel corpus of the two languages under consideration. We propose that the syntactic relationships between the chunks of the input sentence T (of the target language) are given depending upon the relationships of the corresponding chunks in the translation S of T. Along with the parsed structure of the input, the system also outputs the constituent structure (phrases) of the given input sen-

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 301–308,
Sydney, July 2006. ©2006 Association for Computational Linguistics

tence.

In this work, we first discuss the proposed scheme in a general framework. We illustrate the scheme with respect to parsing of Hindi sentences using the Link Grammar (LG) based parser for English and the experimental results are discussed. Before that in the following section we discuss Link Grammar briefly.

2 Link Grammar and Phrases

Link grammar (LG) is a theory of syntax which builds simple relations between pairs of words, rather than constructing constituents in tree-like hierarchy. For example, in an SVO language like English, the verb forms a subject link (`S-`) to some word on its left, and an object link (`O+`) with some word on its right. Nouns make the subject link (`S+`) to some word (verb) on its right, or object link (`O-`) to some word on its left.

The English Link Grammar Parser (Sleator and Temperley, 1991) is a syntactic parser of English based on LG. Given a sentence, the system assigns to it a syntactic structure, which consists of a set of labeled links connecting pairs of words. The parser also produces a "constituent" representation of a sentence (showing noun phrases, verb phrases, etc.). It is a dictionary-based system in which each word in the dictionary is associated with a set of links. Most of the links have some associated suffixes to provide various information (e.g., gender (`m/f`), number (`s/p`)), describing some properties of the underlying word. The English link parser lists total of 107 links. Table 1 gives a list of some important links of English LG along with the information about the words on their left/right and some suffixes.

Link	Word in Left	Word in Right	Suffixes
A	Premodifier	Noun	-
D	Determiners	Nouns	s/m,c/u
J	Preposition	Object of the preposition	s/p
M	Noun	Post-nominal Modifier	p/v/g/a
MV	Verbs/adjectives	Modifying phrase	p/a/i/l/x
O	Transitive verb	Direct or indirect object	s/p
P	Forms of "be"	Complement of "be"	p/v/g/a
PP	Forms of "have"	Past participle	-
S	Subject	Finite verb	s/p, i, g

Table 1: Some English Links and Their Suffixes

As an example, consider the syntactic structure and constituent representation of the sentence given below.

```
       +--------Ss--------+
       |    +----Jp---+   |
 +--Ds-+-Mp-+   +-Dmc-+   +-Pa-+
 |     |    |   |     |   |    |
the teacher of the  boys is   good

    (S (NP (NP The teacher)
           (PP of (NP the boys)))
       (VP is)
       (ADJP good).)
```

It may be noted that in the phrase structure of the above sentence, verb phrase as obtained from the phrase parser has been modified to some extent. The algorithm discussed in this work assumes verb phrases as *the main verb along with all the auxiliary verbs*.

For ease of presentation and understanding, we classify phrase relations as *Inter-Phrase* and *Intra-phrase* relations. Since the phrases are often embedded, different levels of phrase relations are obtained. From the outermost level to the innermost, we call them as "first level", "second level" of relations and so on. One should note that an i^{th} level Intra-phrase relation may become Inter-phrase relation at a higher level.

As an example, consider the parsing and phrase structure of the English sentence given above. In the first level the Inter-phrase relations (corresponding to the phrases "`the teacher of the boys`", "`is`" and "`good`") are `Ss` and `Pa` and the remaining links are Intra-phrase relations. In the second level the only Inter-phrase relationship is `Mp` (connecting "`the teacher`" and "`the boys`"), and the Intra-phrase relations are `Ds`, `Jp` and `Dmc`. In third and the last level, `Jp` is the Inter-phrase relationship and `Dmc` is the Intra-phrase relation (corresponding to "`of`" and "`the boys`").

The algorithm proposed in Section 4 uses pDCA to first establish the relations of the target language corresponding to the first-level Inter-phrase relations of the source language sentence. Then recursively it assigns the relations corresponding to the inner level relations.

3 DCA vis-à-vis pDCA

Direct Correspondence Assumption (DCA) states that the relation between words in source language sentence can be projected as the relations between corresponding words in the (literal) translation in the target language. Direct Projection Algorithm

(DPA), which is based on DCA, is a straightforward projection procedure in which the dependencies in an English sentence are projected to the sentence's translation, using the word-level alignments as a bridge. DPA also uses some monolingual knowledge specific to the projected-to language. This knowledge is applied in the form of Post-Projection transformation.

However with respect to many language pairs syntactic relationships between the words *cannot* always be imitated to project a parse structure from source language to target language. For illustration consider the sentence given in Figure 1. We try to project the links from English to Hindi in Figure 1(a) and Hindi to Bangla in Figure 1(b). For Hindi sentence, links are given as discussed by Goyal and Chatterjee (2005a; 2005b).

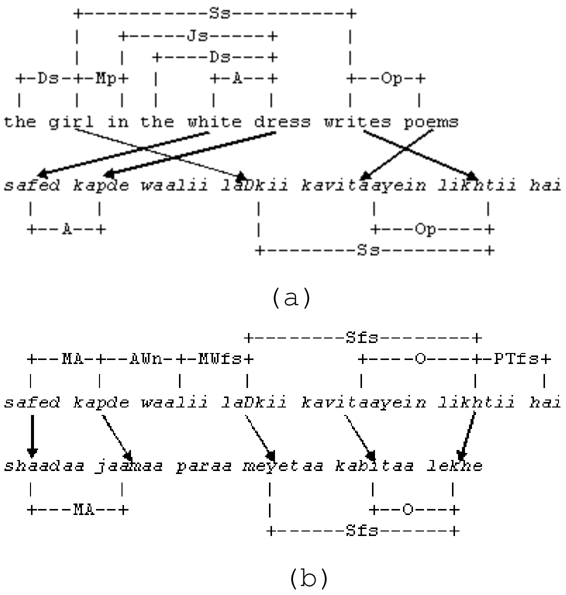

Ss
Js
Ds
Ds
Mp
A
Op
the girl in the white dress writes poems
safed kapde waalii laDkii kavitaayein likhtii hai
A
Op
Ss
(a)
Sfs
MA
AWn
MWfs
O
PTfs
safed kapde waalii laDkii kavitaayein likhtii hai
shaadaa jaamaa paraa meyetaa kabitaa lekhe
MA
O
Sfs
(b)

Figure 1: Failure of DCA

We observe that in the parse structure of the target language sentences, neither all relations are correct nor the parse tree is complete. Thus, we observe that DPA leads to, if not wrong, a very shallow parse structure. Further, Figure 1(b) suggests that DCA fails not only for languages belonging to different families (English-Hindi), but also for languages belonging to the same family (Hindi-Bangla).

Hence it is necessary that the parsing algorithm should be able to differentiate between the links which can be projected directly and the links which cannot. Further it needs to identify the chunks of the target language sentence that cannot be linked even after projecting the links from the source language sentence. Thus we propose pseudo Direct Correspondence Assumption (pDCA) where not all relations can be projected directly. The projection algorithm needs to take care of the following three categories of links:

Category 1: Relationship between two chunks in the source language can be projected to the target language with minor or no changes (for example, subject-verb, object-verb relationships in the above illustration). It may be noted that since except for some suffix differences (due to morphological variations), the relation is same in the source and the target language.

Category 2: Relationship between two chunks in the source language can be projected to the target language with major changes. For example, in the English sentence given in Figure 2(a), the relationship between `the girl` and `in the white dress` is Mp, i.e. "nominal modifier (preposition phrase)". In the corresponding phrases *ladkii* and *safed kapde waalii* of Hindi, although the relationship is same, i.e., "nominal modifier", the type of nominal modifier is changing to *waalaa/waale/waalii*-adjective. If the distinction between the types of nominal modifiers is not maintained, the parsing will be very shallow. Hence the modification in the link is necessary.

Category 3: Relationship between two chunks in the target language is either entirely different or can not be captured from the relationship between the corresponding chunk(s) in the source language. For example, the relationship between the main verb and the auxiliary verb of the Hindi sentence in Figure 2(a) can not be defined using the English parsing. Such phrases should be parsed independently.

The proposed algorithm is based on the above-described concept of pDCA which gives the parse structure of the sentences given in Fig. 2.

While working with Indian languages, we found that outermost Inter-phrase relations usually belong to Category 1, and remaining relations belong to Category 2. Generally an innermost Intra-phrase relation (like verb phrase) belongs to Category 3. Thus, outermost Inter-phrase relations can usually be projected to target language directly, innermost Intra-phrase relations for the target language which are independent of the source language should be decided on the basis of language specific study and remaining relationship should

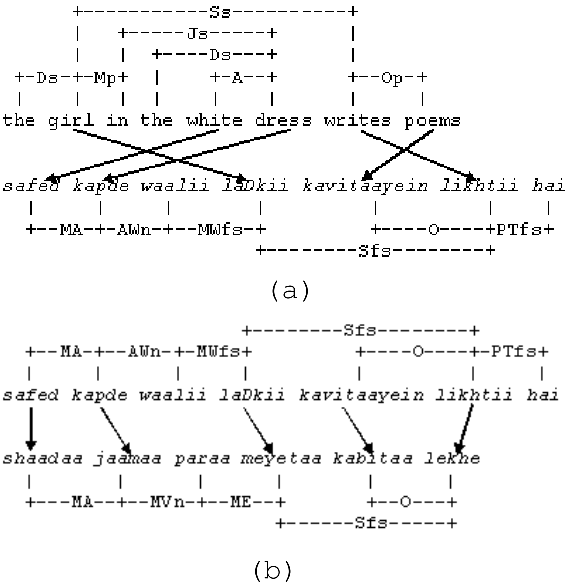

Ss
Js
Ds
Ds
Mp
A
Op
the girl in the white dress writes poems
safed kapde waalii laDkii kavitaayein likhtii hai
MA
AWn
MWfs
O
PTfs
Sfs
(a)
Sfs
MA
AWn
MWfs
O
PTfs
safed kapde waalii laDkii kavitaayein likhtii hai
shaadaa jaamaa paraa meyetaa kabitaa lekhe
MA
MVn
ME
O
Sfs
(b)

Figure 2: Parsing Using pDCA

be modified before projection from source to target language.

4 The Proposed Algorithm

DPA (Hwa et al., 2005) discusses projection procedure for five different cases of word alignment of source-target language: one-to-one, one-to-none, one-to-many, many-to-one and many-to-many. As discussed earlier, DPA is not sufficient for many cases. For example, in case of one-to-many alignment, the proposed solution is to first create a new empty word that is set as head of all multiply aligned words in target language sentence, and then the relation is projected accordingly. But, in such cases, relations between these multiply-aligned words can not be given, and thus the resulting parsing becomes shallow. The proposed algorithm (pDPA) overcomes these shortcomings as well.

The pDPA works in the following way. It recursively identifies the phrases of the target language sentence, and assigns the links between the two phrases/words of the target language sentence by using the links between the corresponding phrases/words in the source language sentence. It may be noted that link between phrases means link between the head words of the corresponding phrases. Assignment of links starts from the outermost level phrases. Syntactic relations between the constituents of the target language phrase(s) for which the syntactic structure does not correspond with the corresponding phrase(s) in the target language are given independently. A list of link rules is maintained which keeps the information about modification(s) required in a link while projecting from the source language to the target language. These rules are limited to closed category words, to parts of speech projected from source language, or to easily enumerated lexical categories.

Figure 3 describes the algorithm. The algorithm takes an input sentence (T) and the parsing and the constituent structure of its parallel sentence (S). Further S and T are assumed to be word-aligned. Initially, S and T are passed to the module ProjectFrom(), which identifies the constituent phrases of S and the relations between them. Then each set of phrases and relations is passed to the module ParseFrom(). ParseFrom() module takes as input two source phrases/words, relation between them, and corresponding target phrases. It projects the corresponding relations in the target language sentence T. ParseFromSpecial() module is required if the relation between phrases of source language can not be projected so directly to the target language. Module Parse() assigns links between the constituent words of the target language phrases $\in \mathscr{P}$. Notations used in the algorithm are as follows:

- By $T' \sim S'$ we mean that T' is aligned with S', T' and S' being some text in the target and source language, respectively.

- Given a language, the head of a phrase is usually defined as the keyword of the phrase. For example, for a verb phrase, the head word is the main verb.

- $\mathscr{P}$ is the exhaustive set of target language phrases for which Intra-phrase relations are independent of the corresponding source language phrases.

- Rule list $\mathscr{R}$ is the list of source-target language specific rules which specifies the modifications in the source language relations to be projected appropriately in the target language.

- Given the parse and constituent structure of a text S, $\Psi_{ij} = \langle S_i,\ S_j,\ L\rangle$, where L is the relation between the constituent phrases/words S_i and S_j of S. $\Psi'_{ij} = \langle T_i,\ T_j\rangle$, $T_i \sim S_i$ and $T_j \sim S_j$. Further, $\Phi_{ij} = \langle \Psi_{ij}, \Psi'_{ij}\rangle$.

```
**ProjectFrom($S', T'$):**    // $S'$ is a source
        // language sentence or phrase, $T' \sim S'$
{
IF $T' \in \mathscr{P}$
THEN    Parse($T'$);
ELSE
    $S' = \{S_1, S_2, \ldots, S_n\}$;    // $S_i$s are
            //constituent phrases/words of $S'$
    $T' = \{T_1, T_2, \ldots, T_n\}$    // $T_i \sim S_i$
    Find all $\Psi_{ij} = \langle S_i, S_j, L\rangle$ from $S'$ and
    corresponding $\Psi'_{ij} = \{T_i, T_j\}$ from $T'$;
    $\Phi_{ij} = \langle \Psi_{ij}, \Psi'_{ij}\rangle$
    For all $i, j$, push ($\mathscr{S}$, $\Phi_{ij}$);
    While !empty($\mathscr{S}$)
        $\Phi$ = pop($\mathscr{S}$);
        IF $L \notin \mathscr{L}$
        THEN    ParseFrom($\Phi$);
        ELSE    ParseFromSpecial($\Phi$);
}
**Parse($T'$):**    // $T'$ is a target language phrase
{
Assign links between constituent words of $T'$
using target language specific rules;
}
**ParseFrom($\Phi$):**    // $\Phi = \langle \Psi, \Psi'\rangle$;
        // $\Psi = \langle S_1, S_2, L\rangle$; $\Psi' = \langle T_1, T_2\rangle$;
{
IF $T_1 \neq \{\phi\}$ & $T_2 \neq \{\phi\}$ THEN
    Find head words $t_1 \in T_1$ and $t_2 \in T_2$;
    Assign relation $L'$ between $t_1$ and $t_2$;    // $L'$
        //is target language link corresponding
        //to $L$ identified using $\mathscr{R}$
IF $T_1$ is a phrase and not already parsed
THEN    ProjectFrom($S_1, T_1$);
IF $T_2$ is a phrase and not already parsed
THEN    ProjectFrom($S_2, T_2$);
}
**ParseFromSpecial($\Phi$):**    // $\Phi = \langle \Psi, \Psi'\rangle$;
        // $\Psi = \langle S_1, S_2, L\rangle$; $\Psi' = \langle T_1, T_2\rangle$;
{
Use target language specific rules to identify if
the relation between $T_1$ and $T_2$ is given by $L'$;
IF true THEN ParseFrom($\Phi$);
ELSE
    Assign required relations using rules;
    IF $T_1$ is a phrase and not already parsed
    THEN    ProjectFrom($S_1, T_1$);
    IF $T_2$ is a phrase and not already parsed
    THEN    ProjectFrom($S_2, T_2$);
}
```

Figure 3: pseudo Direct Projection Algorithm

- $\mathscr{S}$ is a stack of Φ_{ij}s.
- $\mathscr{L}$ is the set of source language relations whose occurrence in parse of some S' may lead to different structure of T', where $T' \sim S'$.

In the following sections we discuss in detail the scheme for parsing Hindi sentences using parse structure of the corresponding English sentence. Along with the parse structure of the input, the phrase structure is also obtained.

5 Case study: English to Hindi

Prior requirements for developing a parsing scheme for the target language using the proposed algorithm are: development of target language links, word alignment technique, phrase identification procedure, creation of rule set $\mathscr{R}$, morphological analysis, development of ParseFromSpecial() module. In this section we discuss these details for adapting a parser for Hindi using English LG based parser.

Hindi Links. Goyal and Chatterjee (2005a; 2005b) have developed links for Hindi Link Grammar along with their suffixes. Some of the Hindi links are briefly discussed in the Table 2. It may be noted that due to the free word order of Hindi, direction can not be specified for some links, i.e., for such links "Word in Left" and "Word in Right" (second and third column of Table 2) shall be read as "Word on one side" and "Word on the other side", respectively.

Link	Word in Left	Word in Right	Directed
S	Subject	Main verb	NO
SN	*ne*	Main verb	NO
O	Object	Main verb	NO
J	noun/pronoun	postposition	YES
MV	verb modifier	Main verb	NO
MA	Adjective	Noun	YES
ME	*aa-e-ii* form of verb	Noun	YES
MW	*waalaa/waale/ waalii*	Noun	YES
PT	*taa-te-tii* form of verb	declension of verb *honaa*	YES
D	Determiner	Head noun	YES

Table 2: Some Hindi Links

Word Alignment. The algorithm requires that the source and target language sentences are word aligned. Some English-Hindi word alignment algorithms have already been developed, e.g.

(Aswani and Gaizauskas, 2005). However, for the current implementation alignment has been done manually with the help of an online English-Hindi dictionary[1].

Identification of Phrases and Head Words.

Verb Phrases. Corresponding to any main verb v_i present in the Hindi sentence, a verb phrase is formed by considering all the auxiliary verbs following it. A list of Hindi auxiliary verbs, along with the linkage requirements has been maintained. This list is used to identify and link verb phrases. Main verb of the verb phrase is considered to be the head word.

Noun and Postposition[2] Phrases. English NP is translated in Hindi as either NP or PP[3]. Also, English PP can be translated as either NP or PP. If the Hindi noun is followed by any postposition, then that postposition is attached with the noun to get a PP. In this case the postposition is considered as the head. Hindi NP corresponding to some English NP is the *maximal span* of the words (in Hindi sentence) aligned with the words in the corresponding English NP. The Hindi noun whose English translation is involved in establishing the Inter-phrase link is the head word. Note that if the last word (noun) in this Hindi NP is followed by any postposition (resulting in some PP), then that postposition is also included in the NP concerned . In this case the postposition is the head of the NP. The system maintains a list of Hindi postpositions to identify Hindi PPs.

For example, consider the translation pair `the lady in the room had cooked the food`∼ *kamre* (`room`) *mein* (`in`) *baiThii huii* (-) *aurat* (`lady`) *ne* (-) *khaanaa* (`food`) *banaayaa* (`cooked`) *thaa* (-).

The phrase structure of the English sentence is `(`NP_1 `(`NP_2 `the lady) (`PP_1 `in (`NP_3 `the room))) (`VP_1 `had cooked) (`NP_4 `the food).`

Here, some of the Hindi phrases are as follows: *kamre mein* and *aurat ne* are identified as Hindi PP corresponding to English PP_1 and NP_2. The words *mein* and *ne* are considered as their head words, respectively. Since the maximal span of translation of words of English NP_1 is *kamre mein baiThii huii aurat* which is followed by postposition *ne*, the Hindi phrase corresponding to NP_1 is *kamre mein baiThii huii aurat ne* with *ne* as the head word. As *huii* and *thii*, which follow the verbs *baiThii*[4] and *banaayaa* respectively, are present in the auxiliary verb list, Hindi VPs are obtained as *baiThii huii* and *banaayaa thaa* (corresponding to VP_1).

Phrase Set $\mathscr{P}$. Hindi verb phrase and postposition phrases are linked independent of the corresponding phrases in the English sentence. Thus, $\mathscr{P} = \{VP, PP\}$.

Rule List $\mathscr{R}$. Below we enlist some of the rules defined for parsing Hindi sentences using the English links (E-links) of the parallel English sentences. Note that these rules are dependent on the target language.

Corresponding to E-link S: If the Hindi subject is followed by *ne*, then the subject makes a `Jn` link with *ne*, and *ne* makes an `SN` link with the verb.

Corresponding to E-link O: If the Hindi object is followed by *ko*, then the object makes a `Jk` link with *ko*, and *ko* makes an `OK` link with the verb.

Corresponding to E-links M, MX: English NPs may have preposition phrase, present participle, past participle or adjective as postnominal modifiers which are translated as prenominal modifiers, or as relative clause in Hindi. The structure of postnominal modifier, however, may not be preserved in the Hindi sentence. If the sentence is not complex, then the corresponding Hindi link may be one of `MA` (adjective), `MP` (postposition phrase), `MT` (present participle), `ME` (past participle), or `MW` (*waalaa/waale/waalii*-adjective). An appropriate link is to be assigned in Hindi sentence after identification of the structure of the nominal modifier. These cases are handled in the module ParseFromSpecial(). The segment of the module that handles English `Mp` link is given in Figure 4.

Further, since morphological information of Hindi words can not be always extracted using corresponding English sentence, a morphological analyzer is required to extract the information[5]. For the current implementation, morphological infor-

[1] www.sanskrit.gde.to/hindi/dict/eng-hin-itrans.html

[2] In Hindi prepositions are used immediately after the noun. Thus, we refer to them as “postposition”.

[3] PP for English is preposition phrase and for Hindi it stands for postposition phrase.

[4] We observe that English PP as postnominal modifier may be translated as verbal prenominal modifier in Hindi and in such cases some unaligned word is effectively a verb.

[5] For Hindi, some work is being carried out in this direction, e.g., http://ccat.sas.upenn.edu/plc/ tamilweb/hindi.html

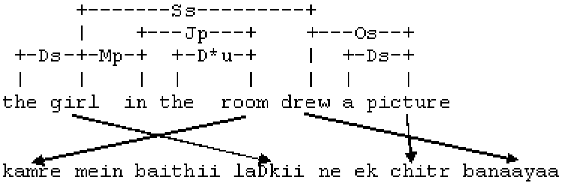
```
ParseFromSpecial($\Phi$):        // $\Phi = \langle \Psi, \Psi' \rangle$;
                                 // $\Psi = \langle S_1, S_2, L \rangle$; $\Psi' = \langle T_1, T_2 \rangle$;
{
IF $L =$ Mp THEN    //$S_1$ and $S_2$ are NP and PP, resp.
    IF $T_2$ is followed by some verb, $v$, not aligned with
    any word in $S$ THEN
        $T_3 =$ VP corresponding to $v$;
        Parse($T_3$);
        Find head word $t_1 \in T_1$;
        Assign MT/ME link between $v$ and $t_1$;
        Assign MVp link between postposition (in $T_2$)
        and $v$;
        ProjectFrom($S_1, T_1$); ProjectFrom($S_2, T_2$);
    ELSE
        ParseFrom($\Phi$);
ELSE
    Check for other cases of $L$;
}
```

Figure 4: ParseFromSpecial() for 'Mp' Link

mation is being extracted using some rules in simpler cases, and manually for more complex cases.

5.1 Illustration with an Example

Consider the English sentence (S) `the girl in the room drew a picture`, its parsed and constituent structure as given in Figure 5. Further, the corresponding Hindi sentence (T), and the word-alignment is also given.



```
      +-------Ss---------+
      |    +---Jp---+    +---Os--+
 +-Ds-+-Mp-+  +-D*u-+    |  +-Ds-+
 |    |    |  |     |    |  |    |
the girl  in the  room drew a picture

kamre mein baithii laDkii ne ek chitr banaayaa
```

```
(S (NP (NP The girl) (PP in (NP the room)))
(VP drew) (NP a picture).)
```

Figure 5: An Example

The step-by-step parsing of the sentence as per the pDPA is given below.

ProjectFrom(S, T):

$S = \{S_1, S_2, S_3\}$, where S_1, S_2, S_3 are the phrases `the girl in the room`, `drew` and `a picture`, respectively. From the definition of Hindi phrases, corresponding T_i's are identified as "*kamre mein baithii laDkii ne*", "*banaayaa*" and "*ek chitr*". From the parse structure of S, Φ's are obtained as $\Phi_{12} = \langle\langle S_1, S_2, \mathtt{Ss}\rangle, \langle T_1, T_2\rangle\rangle$ and $\Phi_{23} = \langle\langle S_2, S_3, \mathtt{Os}\rangle, \langle T_2, T_3\rangle\rangle$. These Φ's are pushed in the stack $\mathscr{S}$ and further processing is done one-by-one for each of them. We show the further process for the Φ_{12}.

Since Ss $\notin \mathscr{L}$, ParseFrom(Φ_{12}) is executed.

ParseFrom(Φ_{12}):

The algorithm identifies t_1 = *ne*, t_2 = *banaayaa*. The Hindi link corresponding to Ss will be SN. The module ProjectFrom(S_1, T_1) is then called.

ProjectFrom(S_1, T_1):

$S_1 = \{S_{11}, S_{12}\}$, where S_{11} and S_{12} are `the girl` and `in the room`, respectively. Corresponding T_{11} and T_{12} are *ladkii ne* and *kamre mein*. Thus, $\Phi = \langle\langle S_{11}, S_{12}, \mathtt{Mp}\rangle, \langle T_{11}, T_{12}\rangle\rangle$. Since L = Mp $\in \mathscr{L}$, ParseFromSpecial(Φ) is called.

ParseFromSpecial(Φ): (Refer to Figure 4)

Since T_2 is followed by an unaligned verb *baithii*, the algorithm finds T_3 as *baithii*, and t_1 as *ne*. It assigns ME link between *baithii* and *ne*. Further, MVp link is assigned between *mein* and *baithii*. Then ProjectFrom(S_{11}, T_{11}) and ProjectFrom(S_{12}, T_{12}) are called. Since both T_{11} and $T_{12} \in \mathscr{S}$, J and Jn links are assigned between constituent words of T_{11} and T_{12}, respectively, using Hindi-specific rules.

Similarly, Φ_{23} is parsed.

The final parse and phrase structure of the sentence are obtained as given in Figure 6.

```
                +-----ME-----+--------SN----+
 +--J--+-MVp-+       +-Jn-+  +-Ds-+--Oms-+
 |     |     |       |    |  |    |      |
kamre mein baiThii laDkii ne ek chitr banaayaa

      (NP(PP (NP kamre) mein) (VP baiThii)
         (PP (NP laDkii) ne))
      (NP ek chitra) (VP banaayaa)
```

Figure 6: Parsing of Example Sentence

6 Experimental Results

Currently the system can handle the following types of phrases in different simple sentences.

Noun Phrase. There can be four basic elements of an English NP[6]: determiner, pre-modifier, noun (essential), post-modifier. The system can handle any combination of the following: adjective, noun, present participle or past participle as pre-modifier, and adjective, present participle, past participle or preposition phrase as post-modifier. Note that some of these cases may be translated as complex sentence in Hindi (e.g., (`book on the table` $\sim$ *jo kitaab mej par rakhii hai*). We are working upon such cases.

[6] Pronouns as NPs are simple.

Verb Phrase. The system can handle all the four aspects (indefinite, continuous, perfect and perfect continuous) for all three tenses. Other cases of VPs (e.g., modals, passives, compound verbs) can be handled easily by just identifying and putting the corresponding auxiliary verbs and their linking requirements in the auxiliary verb list.

Since the system is not fully automated yet, we could not test our system on a large corpus. The system has been tested on about 200 sentences following the specific phrase structures mentioned above. These sentences have been taken randomly from translation books, stories books and advertisement materials. These sentences were manually parsed and a total of 1347 links were obtained. These links were compared with the system's output. Table 3 summarizes the findings.

Correct Links	: 1254
Links with wrong suffix	: 47
Wrong links	: 22
Links missing	: 31

Table 3: Experimental Results

After analyzing the results, we found that

- For some links, suffixes were wrong. This was due to insufficiency of rules identifying morphological information.

- Due to incompleteness of some cases of ParseFromSpecial() module, some wrong links were assigned. Also, some links which should not have been projected, were projected in the Hindi sentence. We are working towards exploring these cases in detail.

- Some links were found missing in the parsing since corresponding sentence structures are yet to be considered in the scheme.

7 Concluding Remarks

The present work focuses on development of Example based parsing scheme for a pair of languages in general, and for English to Hindi in particular.

Although the current work is motivated by (Hwa et al., 2005), the algorithm proposed herein provides a more generalized version of the projection algorithm by making use of some target language specific rules while projecting links. This provide more flexibility in the projection algorithm. The flexibility comes from the fact that unlike DPA the algorithm can project links from the source language to the target language even if the translations are not literal. Use of rules at the projection level gives more robust parsing and reduces the need of post-editing. The proposed scheme should work for other target languages also provided the relevant rules can be identified. Further, since LG can be converted to Dependency Grammar (DG) (Sleator and Temperley, 1991), this work can be easily extended for languages for which DG implementation is available.

At present, we have focused on developing parsing scheme for simple sentences. Work has to be done to parse complex sentences. Once a sizeable parsed corpus is generated, it can be used for developing the parser for a target language using bootstrapping. We are currently working on these lines for developing a Hindi parser.

References

Niraj Aswani and Robert Gaizauskas. 2005. A hybrid approach to aligning sentences and words in English-Hindi parallel corpora. In *ACL 2005 Workshop on Building and Using Parallel Texts: Data-driven machine translation and Beyond*.

Rens Bod, Remko Scha, and Khalil Sima'an, editors. 2003. *Data-Oriented Parsing*. Stanford: CSLI Publications.

Shailly Goyal and Niladri Chatterjee. 2005a. Study of Hindi noun phrase morphology for developing a link grammar based parser. *Language in India*, 5.

Shailly Goyal and Niladri Chatterjee. 2005b. Towards developing a link grammar based parser for Hindi. In *Proc. of Workshop on Morphology*, Bombay.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325, September.

Daniel Sleator and Davy Temperley. 1991. Parsing English with a link grammar. Computer Science technical report CMU-CS-91-196, Carnegie Mellon University, October.

Oliver Streiter. 2002. Abduction, induction and memorizing in corpus-based parsing. In *ESSLLI-2002 Workshop on Machine Learning Approaches in Computational Linguistics,*, Trento, Italy.

David Yarowsky and Grace Ngai. 2001. Inducing multilingual POS taggers and NP bracketers via robust projection across aligned corpora. In *NAACL-2001*, pages 200–207.