Towards Building Parallel Dependency Treebanks: Intra-Chunk Expansion and Alignment for English Dependency Treebank

Debanka Nandi, Maaz Nomani
Jamia Hamdard, New Delhi, India
debanka.nandi0@gmail.com, maaz_nomani@hotmail.com

Himanshu Sharma, Himani Chaudhary, Sambhav Jain, Dipti Misra Sharma
IIIT-H, Hyderabad, India
{himanshu.sharma,himani,sambhav.jain}@research.iiit.ac.in
dipti@iiit.ac.in

Abstract

The paper presents our work on the annotation of intra-chunk dependencies on an English treebank that was previously annotated with Inter-chunk dependencies, and for which there exists a fully expanded parallel Hindi dependency treebank. This provides fully parsed dependency trees for the English treebank. We also report an analysis of the inter-annotator agreement for this chunk expansion task. Further, these fully expanded parallel Hindi and English treebanks were word aligned and an analysis for the task has been given. Issues related to intra-chunk expansion and alignment for the language pair Hindi-English are discussed and guidelines for these tasks have been prepared and released.

1 Introduction

Recent years have seen an increasing interest in research based on parallel corpora. Statistical machine translation systems use parallel text corpora to learn pattern-based rules. These rules can be simple or sophisticated, based on the level of information present in the corresponding parallel corpus. Earlier research in statistical MT utilized just sentence and lexical alignment (Brown et al., 1990) which requires merely a sentence and word aligned parallel text. Later, to acquire these rules the alignment of a parsed structure in one language with a raw string in the other language emerged (Yamada and Knight, 2001; Shen et al., 2008). Of late, the focus has been on exploring these rules from the alignment of source/target language parse trees (Zhang et al., 2008; Cowan, 2008). Also, mapping from a source language tree to a target language tree offers a mechanism to preserve the meaning of the input and producing a target language tree helps to ensure the grammaticality of the output (Cowan, 2008). Thus there is a need for aligned parallel treebanks with alignment information on top of their parsing information.

And, with the availability of a number of treebanks of various languages now, parallel treebanks are being put to use for analysis and further experiments. A parallel treebank comprises syntactically annotated aligned sentences in two or more languages. In addition to this, the trees are aligned on a sub-sentential level. (Tinsley et al., 2009). Further, such resources could be useful for many applications, e.g. as training or evaluation corpora for word/phrase alignment, as also for data driven MT systems and for the automatic induction of transfer rules. (Hearne et al., 2007)

Our work using two parallel dependency treebanks is another such effort in this direction. It includes:

1. Intra-chunk expansion of the English treebank previously annotated with Inter-chunk dependencies, for which there exists a fully expanded parallel Hindi dependency treebank.

2. An analysis of the inter-annotator agreement for the chunk expansion task mentioned in (1) above.

3. Alignment of the two treebanks at sentence and also at word level.

A chunk, by definition, represents a set of adjacent words in a sentence, which are in dependency relation with each other, and where one of these words is their head. (Mannem et al., 2009). The task of dependency annotation is thus divided into: inter-chunk dependency annotation (relations between these chunks) and intra-chunk

Proceedings of the Second International Conference on Dependency Linguistics (DepLing 2013), pages 227–235,
Prague, August 27–30, 2013. © 2013 Charles University in Prague, Matfyzpress, Prague, Czech Republic

dependency annotation (relations between words inside the chunk).

Some notable efforts in this direction include the automatic intra-chunk dependency annotation of an inter-chunk annotated Hindi dependency treebank, wherein they present both, a rule-based and a statistical approach to automatically mark intra-chunk dependencies on an existing Hindi treebank (Kosaraju et al., 2012). Zhou (2008) worked on the expansion of the chunks in the Chinese treebank TCT (Qiang, 2004) through automatic rule acquisition.

The remainder of the paper is organized as follows: In Section 2, we give an overview of the two dependency treebanks used for our work, and their development. Section 3 describes the guidelines for intra-chunk dependency annotation. In Section 4, we talk about issues with the expansion and our resolutions for them. Further, it presents the results of the inter-annotator agreement. Section 5 comprises our work on alignment of parallel Hindi-English corpora and the issues related to that. We conclude and propose some future works towards the end of the paper.

2 Treebanks

We make use of the English dependency treebank (reported in Chaudhry and Sharma (2011)), developed on the Computational Paninian Grammar (CPG) model (Bharati et al., 1995), for this work. This treebank is parallel to a section of the Hindi Dependency treebank (reported in Bhatt et al. (2009)) being developed under the Hindi-Urdu Treebank (HUTB) Project and was created by translating the sentences from HUTB. The English treebank is much smaller in size, with around 1000 sentence (nearly 20K words) as compared to its Hindi counterpart which has about 22800 sentences (nearly 450K words). There is a difference in size of nearly 1000 words between the English treebank and its parallel Hindi treebank from which it was translated. This is because the translations involve both literal and stylistic variations.

The annotation labels used to mark the relations in the treebank conform to the dependency annotation scheme reported in Chaudhry and Sharma (2011), which is an adaptation of the annotation scheme given by Begum et al. (2008), for Hindi. Further, as per these annotation schemes, dependency relations in the treebank are marked at chunk level (between chunk heads), instead of being marked between words.

The Hindi treebank also had intra-chunk dependency relations marked on it, along with the inter-chunk dependencies. And since the English dependency treebank used here, is relatively much newer, there was scope for further work on it. We thus expanded this treebank at intra-chunk level, annotating each node within the chunk with its dependency label/information. Annotating the English treebank with this information brings it at par with the parallel Hindi treebank, making them better suited for experimentation on parallel treebanks.

Further, the earlier version of the treebank was annotated only with the inter-chunk dependencies. Consequently, this enforced a restriction to interpret the chunk merely as a group of words with the head of the group as its representative. The relations among other nodes inside the chunk remained unaccounted for. Now, with the intra-chunk dependencies also marked, the treebank has complete sentence level parsing information, giving access to the syntactic information associated with each node in the tree.

Additionally, the dependency annotation is done using Sanchay annotation interface, and the data is stored in Shakti Standard Format (SSF).[1] (Bharati et al., 2006).

3 Intra-chunk Expansion

As mentioned earlier, the English dependency treebank used here, is relatively much smaller than its parallel Hindi treebank. Given this, we manually expanded this treebank at intra-chunk level, and performed an inter-annotator analysis for this task. Preparation of a set of guidelines for the expansion is another aspect of this effort.

This section of the paper reports our annotation of intra-chunk dependencies (dependency relations among the words within chunks) on the English dependency treebank (described in Section-2) in which inter chunk dependencies are already marked using the CPG model. Adding the intra-chunk annotation provides a fully parsed dependency treebank for English.

The intra-chunk dependencies for this task, were annotated manually (by two annotators). Inter-annotator agreement values for this intra-chunk annotation were then calculated. Both of these tasks are reported in this section, as also, a

[1] http://ltrc.iiit.ac.in/mtpil2012/Data/ssf-guide.pdf

discussion of the types of issues encountered in the annotation.

For the purpose of intra-chunk annotation, the chunk expansion guidelines for the Hindi Treebank expansion were taken as a point of reference and adapted to suit the requirements of the English treebank. The guidelines thus prepared, were used to annotate the intra-chunk dependencies in the English treebank. After the initial annotation and a subsequent analysis of the encountered ambiguous cases, they have been updated accordingly.

The guidelines thus prepared, serve to ensure consistency across multiple annotations. There are a total of 18 intra-chunk tags in the guidelines. The tags are of three types: (a) normal dependencies, eg. nmod__adj, jjmod__intf, etc., (b) local word group dependencies (lwg), eg. lwg__prep, lwg__vaux, etc., and (c) linking part-of dependencies, eg. pof__cn. (Table 1)

Local Word Groups (lwg) are word groups formed on the basis of 'local information' (i.e. information based on adjacent words) (Bharati et al., 1995). 'lwg' dependencies occur due to adjacency of words in a local word group. These are of two types: simple-lwg dependencies and linking-lwg dependencies (termed as 'linking part-of dependencies' above). Linking-lwg dependencies are marked for words that are parts of an LWG, and don't modify each other (usually used in compound nouns, named-entities etc.). Normal dependencies are marked for individual words and don't represent a relation with the complete local word group. For Ex. nmod_adj relation is for a **Noun modifier** of the type **Adjective**. Here the association of the adjective is not with the complete 'lwg', but with a particular noun which may or may not restrict the meaning of the 'lwg'.

4 Inter-Annotator Agreement: Evaluation and Analysis

4.1 Evaluation Criteria

The guidelines for Intra-chunk Expansion were created by studying different possible cases of chunk expansion. The guidelines, in total, list 18 different intra-chunk tags. These intra-chunk labels along with 34 inter-chunk labels make a total of 52 dependency tags. Inter-Annotator Agreement was then calculated on these fully expanded dependency trees for the 2 annotators.

Label Type	Label Description
nmod__adj	Noun modifier of the type adjective
nmod__n	Noun modifier of the type noun
jjmod__intf	Adjective modifier of the type intensifier
lwg__det	A determiner associated with an LWG
lwg__inf	An infinitive marker associated with LWG
lwg__prep	A preposition associated with LWG
lwg__neg	A negation particle associated with LWG
lwg__vaux	An auxiliary verb associated with LWG
lwg__rp	A particle associated with LWG
lwg__uh	An interjection particle associated with LWG
lwg__poss	A possession marker associated with LWG
lwg__adv	An adverb associated with LWG
lwg__ccof	Arguments of a conjunct associated with LWG
lwg__emph	An emphatic marker associated with LWG
lwg__v	Verbal nouns (participials, gerunds etc.) associated with LWG
pof__cn	Part-of relation expressing continuation
pof__redup	Part-of relation expressing reduplication
rsym	Symbols

Table 1: Label Types and Descriptions

Fleiss's Kappa (Fleiss, 1971) is used to calculate the agreement, which is a commonly used measure for calculating agreement over multiple annotators. Table 3 shows the agreement strength relative to the kappa statistic. The Fleiss's kappa is calculated as :

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

The factor 1 - Pr(e) gives the degree of agreement that is attainable above chance, and, Pr(a) - Pr(e) gives the degree of agreement actually achieved above chance.

$$Pr(a) = \frac{1}{Nn(n-1)} \sum_{i \in N} \sum_{j \in k} (n_{ij}^2 - Nn)$$

$$Pr(e) = \sum_{j \in k} p_j^2 \qquad \text{where,}$$

$$p_j = \frac{1}{Nn} \sum_{i \in N} n_{ij}$$

Along with the Fleiss Kappa, we also calculate the Unlabelled Attachment and Labelled Attachment accuracies for the fully expanded trees in Table 4. The inter-chunk labels were not excluded for calculating the above mentioned statistics. This is because identifying the head in a chunk is also an important step in creating a fully connected tree. It has been further analysed in Section 4.2 and shown that identifying a different head might lead to different fully expanded trees, and therefore, must be included in the calculation of final statistics.

Edge Pairs	Unlabelled Attachment (UA)	Label Accuracy	Labelled Attachment
1718	1605 (93.42%)	1611 (93.77%)	1554 (90.45%)

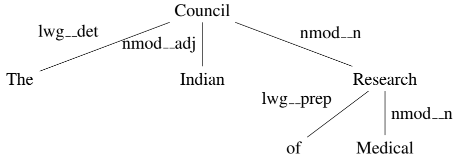
Table 4: Attachment and Label Accuracy

Edge Pairs	Agreement	Pr(a)	Pr(e)	Kappa
1605	1554	0.955	0.061	0.952

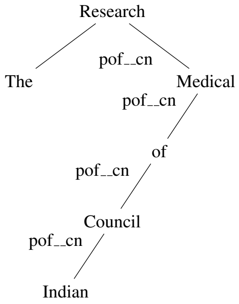
Table 2: Kappa statistics for Inter-Annotator Experiment

Kappa Statistic	**Strength of agreement**
<0.00	Poor
0.0-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Table 3: Coefficients for the agreement-rate based on (Landis and Koch, 1977).



S1 : The Indian Council of Medical Research (Chunk Analysis)*



S2 : The Indian Council of Medical Research (Frozen)

4.2 Analysis

Besides calculating the inter-annotator agreement, cases with disagreement were analysed for possible errors and cases of ambiguities in the guidelines. The observed cases which led to percentage error in inter-annotator agreement were then resolved and the guidelines were updated, so as to reduce potential errors arising due to these in future. A few of the observed cases have been discussed below:

1. **Named Entity Handling**

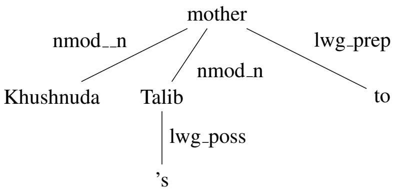
   Since the treebank doesn't have Named Entity (NE) annotation, the handling of NEs induced an element of disagreement between the two annotations. Ex. In the case below, **The Indian Council of Medical Research** is an NE, but it has been handled differently by the two annotators.

   Whether it should be treated as a frozen unit (A compound noun with no further analysis in the structure of the name), or it should be treated as a phrase that is analysed for the association of constituents is the issue here.

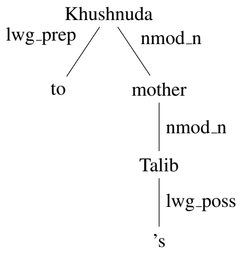
   We chose to consider structure-2 as the appropriate one. This decision is motivated by the observation that Named Entities are frozen expressions and may or may not always be analysable in parts. This will thus help maintain consistency in annotation of NEs across the treebank.

2. **Appositives**

   Appositives are grammatical constructions where two noun-phrases are placed adjacent to each other and one modifies or restricts the other. In the PP phrase below, **to Talib's mother Khushnuda**, there are two noun phrases **Talib's mother** and **Khushnuda** (name) and any of them can be considered as the head of the chunk. Further, the preposition **to** can attach to any of the two noun phrases.



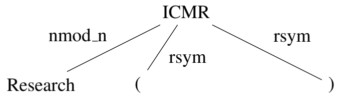
S1 : to Talib's mother Khushnuda*



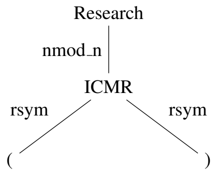
S2 : to Talib's mother Khushnuda

For these cases, the noun phrase that most specifically describes the object of discussion is taken to be the primary noun phrase and the secondary ones as its modifiers. For the above example, **Khushnuda** is the NP that specifies the head of the phrase more clearly and is thus considered to be the head, and the preposition **to** is attached to **Khushnuda**, rendering *S2* as the correct analysis.

In cases of abbreviations, where both noun phrases are different representations of the same name, we consider the expanded name to be the head and the abbreviation is attached as a modifier of the head noun. In the example below, since **The Indian Council of Medical Research** is being considered a Named Entity, **Research** is the head of the phrase and the abbreviation **ICMR** is attached to it as a modifier.



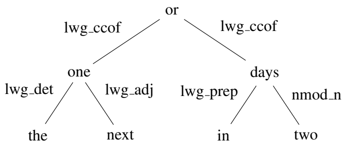
S1 : The Indian Council of Medical Research(ICMR)*



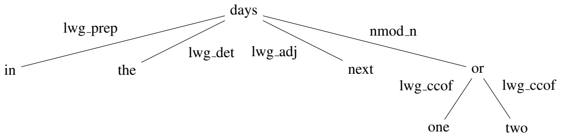
S2 : The Indian Council of Medical Research(ICMR)

3. **Head-Identification**

While annotating relations between tokens, identifying the head of a constituent is a crucial step and decides the structure of the fully expanded tree. Ex. in the PP phrase **in the next one or two days**, the most probable head is **days**. In our scheme, the coordinator is considered to be the head of the coordinated phrase, hence **or** is regarded as the head of **one** and **two** (*S1*). Another possibility is to add a NULL element in the first argument of conjunction and make the phrase **in the next one NULL or two days**, where the **NULL** is a copy of the features of **days** (*S2*).



S1 : in the next one or two days*

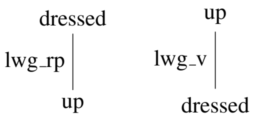


S2 : in the next one or two days

However, in the inter-chunk dependency annotation scheme NULLs are inserted only if they are crucial for representing the dependency structure. Following this, *S2* was preferred over *S1* for such cases. Also, in *S2* the association of cardinals **one** and **two** with **days** is easily visible and can be interpreted if required.

4. **Phrasal Verbs**

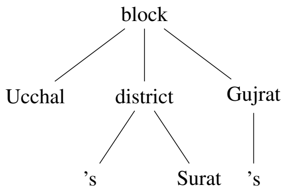
Phrasal verbs are verbs that include particles or prepositions. Their meaning is non-compositional, as it cannot be retrieved by individually handling the lexical items. Ex. **look after** : verb+preposition , **brought up** : verb+particle , **put up with** : verb+particle+preposition . For these cases, the verb is considered to be the head of the chunk. A clear distinction between prepositions and particles in a verb phrase has been made in our guidelines by way of different annotation labels. The associated labels are: *lwg_prep* (local-word-group preposition) and *lwg_rp* (local-word-group particle). A few examples of this are:



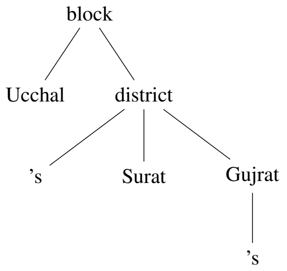
dressed up *dressed up**

5. **Genitives**

We observed disagreement between the two annotations where there were instances of multiple consecutive genitives in a single chunk. However, this cannot be resolved at the level of guidelines since the decisions in such cases would depend on the world knowledge of annotators and would have to be resolved contextually and individually. The example below illustrates this further.



S1 : Gujrat's Surat district's Ucchal block*



S2 : Gujrat's Surat district's Ucchal block

Here the knowledge whether **Surat** is a **district** in **Gujrat** or not is important in deciding if **Gujrat** should modify **District** (in *S2*) or **block** (in *S1*). Here, since **Surat** is a district in **Gujrat**, **Gujrat**, *S2* would be the correct analysis rejected.

5 Alignment

In this task, the fully expanded English dependency trees, obtained after the intra-chunk expansion, were aligned with their respective counterparts in the Hindi Dependency Treebank(Fully expanded Hindi dependency trees).

Due to limitations of the available annotation tools, one cannot align trees from one language to the other directly in a structural manner. Thus, we chose to align the data at the textual level and then incorporated them in the already existing treebank. *Sanchay*[2] was chosen as the alignment tool after experimenting with some openly available tools such as GATE, Cairo etc.

The alignment was done in two stages :

1. *Sentence Alignment* : First, parallel text files (Hindi and English) were aligned on the sentence level.
2. *Word Alignment* : A set of guidelines were created for word alignment, by doing a pilot study on a small dataset. As we encountered issues during the alignment, these guidelines were updated accordingly.

After the two alignment tasks, the word aligned data was merged with the respective treebanks.

5.1 Issues in Alignment

5.1.1 Sentence Alignment Issues

For sentence alignment, a basic postulate was that all the events must be captured in a sentence aligned pair [(source sentences)-(target sentences)]. As is commonly observed in studies of parallel corpora, the target language sometimes removes argument information, or adds extra arguments to provide a sound translation. These cases are not considered as a divergence at the level of sentence alignment, since the selection criteria is strictly limited to event information. In our task, we encountered 4 types of sentence alignment structures. These are :

1. *One-to-One Mapping* : When all the events in a source sentence are mapped to events in the target sentence, we say that there is a One-to-One mapping. For instance, in Figure-1, all the source sentences are mapped to exactly one target sentence, although crossed.

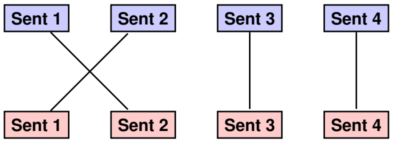


Figure 1: One-to-One Alignment

2. *Many-to-One Mapping* : Cases where multiple source sentences map to a single sentence in the target language, i.e. events in multiple source sentences are incorporated into a single target sentence. In Figure-2, Sent-2, Sent-3 and Sent-4 of the source language go to Sent-3 of the target Language.
3. *One to Many Mapping* : Single source language sentence is divided into multiple target language sentences. In Figure-2, Sent-1 of source language is aligned to both Sent-1 and Sent-2 of the target language.

[2] Sanchay : http://www.sanchay.co.in

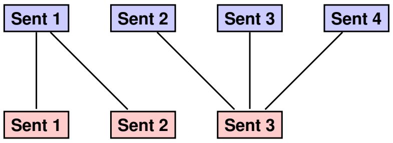


Figure 2: One-to-Many, Many-to-One Alignment

4. *Many to Many Mapping* : Events are distributed unevenly in source and target sentences. In example x, for a pair of source and target sentences, the mapping resembles a 'Z' structure. Such sentences were altered to convert them into one of the above three types. Figure-3 shows the Z-structure observed in those cases. This particular case can be resolved by changing the sentence alignments as per Figure-4.

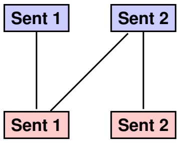


Figure 3: Many-to-Many Alignment

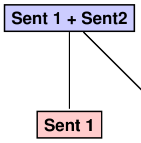


Figure 4: Many-to-Many Alignment(Altered)

5.1.2 Word Alignment Issues

During word alignment, the main focus was on the maintaining the syntactic and semantic functions of the words across the language pair. For many cases, it was not possible to syntactically align the words, as is observed in stylistic translations, idiomatic usages, multi-word expressions etc. The similarity in the semantic function of the words was the deciding factor for these cases.

During the course of annotation and while developing the guidelines, few issues related to word alignment were encountered. Given below is a summary of the types of divergences that were found, with a few examples.

1. **Multi Word Expressions(MWE)** :
   Multi word expressions in source languages are translated to another MWE in the target language or vice-versa. These are divided into two types :
   - *Many-to-One* OR *One-to-Many* alignments are the those where MWE in one language maps to a single word in another language. In such cases, all the constituting words of the MWE are mapped to that single word in the target language. Ex. Hindi : '***aguvAI karne vAle***' goes to English '***heading***' in Example-1.

     (1) .. *xala* *kI* ***aguvAI*** ***karne***
     .. group of head do
     vale *KAna* ..
     ATTR Khan ..
     ".. Khan **heading** the group .."
   - In cases of *Many-to-Many* alignment, where the MWE is literally translated with/without retaining the sense, we map each individual token of the MWE to their respective mappings in the other language. Thus, essentially reducing the problem to either *One to One*, *One to Many* or *Many to One*. For cases, where MWEs are not literally translated, we map all the tokens in the source language MWE to the head of the MWE chunk in the target language.

2. **Mismatched syntactic categories** :
   Many syntactic categories like determiners, infinitives are realized differently (syntactically/structurally) in Hindi. Ex. English determiner '***A***' goes to Hindi ordinal '***eka***' sometimes, and doesn't have a mapping for other cases . These functional categories are mapped to either the token aligned with the head of their chunk (category) or to the element which is functionally similar. In Example-2, English determiner '***every***' is mapped to Hindi noun '***kaxama***'. In this particular case, Hindi employs the use of reduplication to get the same meaning as the determiner '***every***'.

   (2) .. ***kaxama*** ***kaxama*** *para*
   .. step step at
   BraStACAra *hE* ..
   corruption is ..
   ".. corruption is at **every step** .."

Tag	RP	CC	NEG	J	N	QF+	QC	DEM	RB	V	PSP	PRP
FW	0	3	0	0	4	0	0	0	0	0	0	0
V	10	3	11	133	287	4	2	2	7	1323	114	8
PRP+	4	2	0	6	21	0	0	8	0	8	10	237
DT	6	4	8	73	445	24	40	93	5	9	24	55
RP	0	0	0	2	7	0	0	0	2	17	5	0
NNC	0	0	0	0	5	0	0	0	0	0	2	0
TO	2	3	0	6	8	0	0	0	2	37	123	2
RB+	68	24	52	20	48	10	5	3	29	14	36	50
CC	2	163	0	0	2	0	0	2	2	0	7	2
J	4	2	5	229	104	24	11	5	9	28	34	8
N	3	0	7	80	2457	13	16	16	7	44	161	20
IN	14	132	5	10	115	8	5	6	10	49	661	27
CD	2	0	0	3	18	0	82	0	0	2	0	0
MD	4	0	4	8	3	0	0	0	2	85	3	0

Table 5: POS Mappings

3. **Syntactic difference** :
   In cases where a certain word/phrase is present in the source language, while its equivalent is absent in the target. These differences arise due to many reasons including stylistic variation, syntactic differences (word-order), Many to One MWE mappings etc. For Ex. Post-positions in Hindi have a certain mapping to prepositions in English. Though, the prepositions don't always realize. For Ex. *Hindi chunk: rAma ne* is aligned to *English chunk: Ram*. Here, there is no preposition to align with the post-position *ne*. In such cases, the dependents are attached to the word in the target language aligned to it's head element. Thus, the post-position *ne*, here, is aligned to *Ram*.

   It may be noted that, this issue is different from (2) (Mismatched Syntactic Categories) where the meaning of the phrase was being realized in the sentence via some other word that belongs to a category different from the category of the source-word. Here, it is not possible to align individual words due to position, word-order, nature of MWE (literal/metaphorical) and other issues which arise due to syntactic differences between the two languages.

5.1.3 POS Tag

For the purpose of analysis and manual alignment quality evaluation, the POS tag mappings were recorded in a table (Table-5). The POS tagsets are different for English and Hindi treebanks. The English Treebank uses Penn POS Tagset (Marcus et al., 1993), while Hindi treebank is annotated as per Bharati et al. (2006). Keeping the large number of POS categories and differences in tagsets in view, some POS category columns have been merged in the table. For Ex. We merged the categories for JJ,JJR,JJS into J (Adjective) for English POS tagset and JJC,JJ into J for Hindi POS tagset for comparison and error analysis over broad syntactic categories. Major categories such as verbs, adjectives, adverbs, nouns etc. have a considerable mapping ratio in the word aligned data. All the odd POS alignment pairs, such as PostPosition-Determiner, Verb-Preposition, Question Words-Prepositions and many more, were studied and wherever deemed possible, errors in POS tags of these cases were corrected. Cases related to the above-mentioned issues were documented and will be available with the parallel treebank.

Conclusion and Future Work

In this paper we reported our work on Intra-Chunk Annotation and Expansion of the English Treebank, inter-annotator studies over the same, and furthermore, alignment over the expanded parallel data. The reported inter-annotator reliability measure value for intra-chunk expansion was $\kappa = 0.95$. A further analysis of the ambiguous cases was done and the guidelines were further improved so as to resolve the cases of confusion. We extended this work with alignment over parallel Hindi-English fully expanded dependency treebanks in the CPG formalism. A set of guidelines are also prepared for manual alignment of data in Hindi-English language pair. The POS Matrix analysis could provide some insights in the divergences between the two languages. This work could prove helpful in bi-text projections, language divergence studies and statistical machine translation and we hope to take these as our future work.

Acknowledgments

We gratefully acknowledge the provision of the useful resource by way of the Hindi Treebank developed under HUTB, of which the Hindi treebank used for our research purpose is a part, and the work for which is supported by the NSF grant (Award Number: CNS 0751202; CFDA Number: 47.070). Also, any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

R. Begum, S. Husain, A. Dhwaj, D.M. Sharma, L. Bai, and R. Sangal. 2008. Dependency annotation scheme for indian languages. In *Proceedings of IJCNLP*.

Akshar Bharati, Rajeev Sangal, and Vineet Chaitanya. 1995. *Natural Language Processing: A Paninian Perspective*. Prentice-Hall of India.

Akshar Bharati, D.M. Sharma, Lakshmi Bai, and Rajeev Sangal. 2006. Anncorra: Guidelines for pos and chunk annotation for indian languages. Technical report, IIIT-H.

R. Bhatt, B. Narasimhan, M. Palmer, O. Rambow, D.M. Sharma, and F. Xia. 2009. A multi-representational and multi-layered treebank for hindi/urdu. In *Proceedings of the Third Linguistic Annotation Workshop*, pages 186–189. Association for Computational Linguistics.

P.F. Brown, J. Cocke, S.A.D. Pietra, V.J.D. Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

H. Chaudhry and D.M. Sharma. 2011. Annotation and issues in building an english dependency treebank.

B.A. Cowan. 2008. *A tree-to-tree model for statistical machine translation*. Ph.D. thesis, Massachusetts Institute of Technology.

J.L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

M. Hearne, J. Tinsley, V. Zhechev, and A. Way. 2007. Capturing translational divergences with a statistical tree-to-tree aligner.

P. Kosaraju, B.R. Ambati, S. Husain, Sharma, D.M., and R. Sangal. 2012. Intra-chunk dependency annotation: Expanding hindi inter-chunk annotated treebank.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

P. Mannem, H. Chaudhry, and A. Bharati. 2009. Insights into non-projectivity in hindi. In *Proceedings of the ACL-IJCNLP 2009 Student Research Workshop*, pages 10–17. Association for Computational Linguistics.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330.

Z. Qiang. 2004. Annotation scheme for chinese treebank. *Journal of Chinese Information Processing*, 18(4):1–8.

L. Shen, J. Xu, and R. Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. *Proceedings of ACL-08: HLT*, pages 577–585.

J. Tinsley, M. Hearne, and A. Way. 2009. Exploiting parallel treebanks to improve phrase-based statistical machine translation. *Computational Linguistics and Intelligent Text Processing*, pages 318–331.

K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 523–530. Association for Computational Linguistics.

M. Zhang, H. Jiang, A. Aw, H. Li, C.L. Tan, and S. Li. 2008. A tree sequence alignment-based tree-to-tree translation model. *Proceedings of ACL-08: HLT*, pages 559–567.

Q. Zhou. 2008. Automatic rule acquisition for chinese intra-chunk relations. In *Proceedings of International Joint Conference of Natural Language Processing (IJCNLP)*.