# Rapid Porting of DUSTer to Hindi

BONNIE J. DORR
Department of Computer Science and UMIACS, University of Maryland, College Park
NECIP FAZIL AYAN, NIZAR HABASH, NITIN MADNANI
Institute for Advanced Computer Studies, University of Maryland, College Park
REBECCA HWA
Department of Computer Science, University of Pittsburgh, Pittsburgh

The frequent occurrence of *divergences*—structural differences between languages—presents a great challenge for statistical word-level alignment and machine translation. This paper describes the adaptation of DUSTer, a divergence unraveling package, to Hindi during the DARPA TIDES-2003 Surprise Language Exercise. We show that it is possible to port DUSTer to Hindi in under 3 days.

Categories and Subject Descriptors: 1.2.7 [**Artificial Intelligence**]: NLP—Machine Translation

## 1. INTRODUCTION

Word-level bilingual alignments are an integral part of statistical machine translation models. The frequent occurrence of *divergences*—structural differences between languages—presents a great challenge to the alignment task. This paper describes the adaptation of DUSTer (Divergence Unraveling for Statistical Translation) [Dorr et al. 2002] to Hindi during the DARPA TIDES-2003 Surprise Language Exercise.[1] DUSTer is a method for systematically identifying common divergence types and transforming an English sentence structure to bear a closer resemblance to that of another language (henceforth referred to as the *foreign* language). Our goal is to enable more accurate alignment and projection of dependency trees in another language without requiring any training on dependency-tree data in that language. The input text is parsed on the English side only.[2] The projected foreign-language trees may serve as input for training parsers in a new language. We evaluate the usefulness of our approach in terms of the time that it took to complete the process

[1] For more details, see http://www.umiacs.umd.edu/labs/CLIP/DUSTer/Surprise.html.
[2] This work contrasts that of [Gupta and Chatterjee 2003] and [Ding et al. 2003], where parsing is required on both sides.

Authors' addresses: Bonnie J. Dorr, Necip Fazil Ayan, Nizar Habash, Nitin Madnani, Dept of Computer Science and UMIACS, A.V.W. Building, University of Maryland, College Park, MD, 20742; Rebecca Hwa, Dept of Computer Science, University of Pittsburgh, Pittsburgh, PA, 15260.

of porting DUSTer to Hindi.

Consider the case of a manner-of-motion divergence where a verb in one language is expressed as two independent verbs in another language, e.g., the English phrase *run into the room* and its Spanish equivalent *entrar en el cuarto corriendo* (*enter into the room running*). While seemingly transparent for human readers, the frequent occurrence of divergences of this type throws statistical aligners for a serious loop. For example, a good automatic alignment system will be able to detect that *run* maps to *running* but it will leave *enter* unaligned.[3] A preliminary investigation on a sample size of 19K sentences from the TREC El Norte Newspaper (Spanish) Corpus[4] reveals that divergences of this type occurred in approximately 1 out of every 3 sentences.[5] Thus, finding a way to deal effectively with these divergences and repair them would be a massive advance for bilingual alignment.

DUSTer provides a method for automatic detection and processing of divergences, enabling improved alignment and construction of a noise-reduced dependency treebank for training foreign-language parsers.[6] The approach involves transformation of an English dependency tree (produced either by Minipar [Lin 1998] or the Collins parser [Collins 1996]) into a pseudo-English form, $E'$. This form is intended to be more closely matched to the surface form of the foreign language, e.g., "run into the room" is transformed to a form that roughly corresponds to "move in the room running" if the foreign language is Spanish. This rewriting of the English sentence increases the likelihood of one-to-one correspondences, thus facilitating statistical alignment. Given a corpus, divergences are identified, rewritten, and then run through the statistical aligner of choice (e.g., ProAlign [Lin and Cherry 2003] or Giza++ [Al-Onaizan et al. 1999; Och and Ney 2000]).

## 2. PORTING OF DUSTER TO HINDI

Prior to the Hindi Surprise Language exercise, we investigated divergences in several large-scale multilingual corpora (Spanish, Arabic and Chinese). Our investigation revealed that there are 6 divergence types of interest. Once the Hindi exercise began, we used the surprise-language data (e.g., BBC, EMILE, and the electronic Bible) to fill out these divergence types with Hindi examples. Table I shows examples of each type from our corpora, along with examples of sentence pairs.

We accommodate divergence cases through the application of pre-stored divergence transformations—117 "universal rules." The native speaker deemed 21 rules to be applicable to Hindi; no additional universal rules were needed.[7] Table II shows a sample of the universal rules. Each rule has a left-hand side (corresponding to the English string) and a right-hand side (corresponding to the foreign-language string). The rules fall into two categories: Type I rules facilitate the task of align-

[3]Similar cases are discussed in [Lin and Cherry 2003].

[4]LDC catalog no LDC2000T51, ISBN 1-58563-177-9, 2000

[5]This was detected by using automatic detection techniques followed by human confirmation [Dorr et al. 2002].

[6]A similar approach is that of [Carbonell et al. 2002] in which translation rules are learned from an elicited, human-aligned bilingual corpus. The DUSTer approach is different in that the rules are constrained according to human-specified parameter settings, or lexical triggers.

[7]See [Dorr et al. 2002] for a discussion of a corpus-based justification of the divergence classes for several seemingly diverse language pairs, including English-Spanish and English-Arabic.

Table I. Examples of True English, $E'$, and Foreign Equivalent

| Type | English | E' | Foreign Equivalent |
|---|---|---|---|
| Light Verb | **make cuttings** | **wound** | H: काटा |
| Manner | the land **mourns** | the land **stays mourning** | H: धरती रोती रहती है |
| Structural | envy **him** | envy **PREP him** | H: उस से जलता हूँ |
| Categorial | I am **afraid** | to-me **fear** be | H: मुझे डर है |
| Head-Swapping | **is valued** at 4 rupees | **value be** 4 rupees | H: मूल्य चार रुपये है |
| Thematic | **I** am **pained** | **to-me** pain **they** | H: मुझे दुख देते हैं |

Table II. Transformation Rules between E and $E'$

**Type I. Rules Impacting Alignment and Projection**

(1) **Light Verb**

1A. Expansion: $[V_i(\text{PsychV})\ \text{Arg}_j] \rightarrow [V(\text{LightVB})\ \text{Arg}_j\ N_i]$

Ex: "I fear" → "I have fear"

1B. Contraction: $[V(\text{LightVB})\ \text{Arg}_i\ \text{Adj}_j] \rightarrow [V_j(\text{DirectionV})\ \text{Arg}_i]$

Ex: "our hand is high" → "our hand heightened"

(2) **Manner**

2A. Expansion: $[V_i\ \text{Arg}_j] \rightarrow [V(\text{MotionV})\ \text{Arg}_j\ V_i]$

Ex: "I teach" → "I walk teaching"

2B. Contraction: $[V(\text{ChangeOfStateV})\ \text{Arg}_i\ \text{Modifier}_j] \rightarrow [V_j(\text{DirectionV})\ \text{Arg}_i]$

Ex: "he turns again" → "He returns"

(3) **Structural**

3A. Expansion: $[V_i\ \text{Arg}_j\ \text{Arg}_k] \rightarrow [V_i\ \text{Arg}_j\ P(\text{Oblique})\ \text{Arg}_k]$

Ex: "I forsake thee" → "I forsake of thee"

3B. Contraction: $[V_i\ \text{Arg}_j\ P(\text{Oblique})\ \text{Arg}_k] \rightarrow [V_i\ \text{Arg}_k\ \text{Arg}_k]$

Ex: "I search for him" → "I search him"

**Type II. Rules Impacting Projection Only**

(4) **Categorial**

$[V(\text{LightVB})_i\ \text{Arg}_j\ \text{Adj}(\text{Arg}_k)] \rightarrow [V(\text{LightVB})_i\ \text{Arg}_j\ N(\text{Arg}_k)]$

Ex: "I am jealous" → "I have jealousy"

(5) **Head-Swapping**

$[V_i(\text{MotionV})\ \text{Arg}_j\ P_k(\text{DirectionP})] \rightarrow [V_k(\text{DirectionV})\ \text{Arg}_j\ V_i(\text{MotionV})]$

Ex: "I run in" → "I enter running"

(6) **Thematic**

$[V_i\ \text{Arg}_j\ \text{Arg}_k] \rightarrow [V_i\ \text{Arg}_j\ P(\text{Oblique})\ \text{Arg}_k]$

Ex: "He wears it" → "It is-on him"

ment *and* enable more accurate projection of dependency trees. Type II rules *only* enable more accurate projection of dependency trees with minimal or no change to alignment accuracy. In the first category, rules are sub-divided into: (A) *expansion rules*, which are applied when the foreign language sentence is verbose relative to the English one; and (B) *contraction rules* which are applied when the foreign language sentence is terse relative to English.[8]

All E-to-$E'$ universal rules are *parameterized* according to the requirements of the left- and right-hand languages. Parameters are indicated by parenthesized labels, e.g., "LightVB." These correspond to a set of lexical items that are pre-specified by a native speaker of each language, serving as lexical triggers for rule applicability. For example, the verb on the right-hand (Hindi) side of rule 1A is associated with the *LightVB* parameter, which means the choice of possible instantiations of this verb is limited to the human-specified words होना (be), करना (do), बनाना (make), लगना (give), लेना (take), and डालना (put). Similarly, the Verb on the left-hand (English) side of rule 1B is associated with the *LightVB* parameter. This means

[8] We found the expansion rules applied more frequently than contraction rules to Hindi and Spanish, both verbose relative to English, as opposed to the more terse Arabic.

Table III. Times for Human Porting of DUSTer: Hindi, Arabic and Chinese

| Task | Hindi | Arabic | Chinese |
|---|---|---|---|
| **Parameter Setting** | 3.7 hours ≡ 0.5 days | 3.3 hours ≡ 0.4 days | 17.15 hours ≡ 2.1 days |
| **Morph Specification** | 8 hours = 1 day | 16 hours = 2 days | 0 hours = 0 days |
| **Total Time** | 1.5 days | 2.4 days | 2.1 days |

the choice of possible instantiations of this verb is limited to the human specified words *be, do, give, have, make, take,* and *put*.

The rapid setting of these parameters facilitates the porting of DUSTer to new languages. This process involves human translation of 16 English parameter settings to their foreign-language counterparts.[9] Table III indicates the amount of time it took to set the Hindi parameters. The entire porting process took under 3 person-days for Hindi. For comparison, we show the time it took to develop settings for Arabic and Chinese. As in the case of Hindi, the porting process took well under 3 person-days by a native speaker of each of these languages.

The Hindi speaker reported the most difficulty with translating the MotionV and ChangeOfStateV parameters settings to Hindi because of the ambiguity of the original English terms. In addition, determining the applicability of the 117 rules was a difficult task that might be facilitated by a visualization tool (e.g., that of [Carbonell et al. 2002]) in the future.

It is interesting to note that, although the parameter-setting task was *shorter* for Hindi and Arabic than for Chinese, the task of producing morphological variants for each word in the parameters was *longer* for these two languages because of their morphologic richness. The specification of morphological variants took 2 days for Arabic (very rich morphology), 1 day for Hindi (less rich morphology), and no time for Chinese (essentially no morphology). Thus, there is a time tradeoff that balances out all three languages in the end: the overall time for incorporating a new language into DUSTer (i.e., parameter setting plus adding morphological variants) comes out to be about the same for all three languages: 1.5–2.5 days.

We ran DUSTer on 195 Hindi-English sentence pairs from the parallel corpora provided in the surprise-language exercise. The DUSTer run took one hour to produce these results. For example, in the English-Hindi case of *The book was valued at 600 rupees*, DUSTer transforms the English dependency tree into a new dependency tree corresponding to the sentence *The book value was 600 rupees*. The output is: 'The book value(Noun) LightVB 600 rupees'. With this rewritten $E'$ string, we can produce more accurate alignments than would otherwise be possible with direct alignments. Ultimately, these more accurate alignments provide support for improved projection of English dependency trees to Hindi.

## 3. CONCLUSIONS AND FUTURE WORK

We have shown that it is possible to port DUSTer to Hindi in under 3 days, by virtue of setting a small number of parameters for 117 universal mapping rules.

[9] The parameters are: AspectV, ChangeOfStateV, Complement, DirectionP, DirectionV, FunctionalDet, FunctionalN, LightVB, LocationV, ModalV, MotionV, Neg, Oblique, Pleonastic, PsychV, and TenseV.

Currently we are running human alignment experiments on the 195 Hindi-English sentences to determine whether the English/Hindi alignments induced from the automatic E′/Hindi alignments are more accurate than the English/Hindi alignments produced directly. Our measure of accuracy will be based on the degree to which the induced and direct automatic alignment results match those of the human subjects on the 195 Hindi-English sentence pairs. We also plan to evaluate the effect of divergence handling on the foreign parse trees. Our current experiments involve projection of English trees to Hindi. We will compare our approach to alternative dependency-tree projection approaches, e.g., [Hwa et al. 2002].

## ACKNOWLEDGMENTS

This work has been supported in part by DARPA TIDES Cooperative Agreement N66001-00-2-8910, Army Research Lab Cooperative Agreement DAAD190320020, ONR MURI Contract FCPO.810548265, and NSF CISE Research Infrastructure Award EIA0130422. We are grateful for the design and programming expertise of Andrew Fister, Eric Nichols, and Lisa Pearl and to Tiejun Zhao for providing us with Chinese parameter settings. We also thank our Spanish aligners, Irma Amenero, Emily Ashcraft, Allison Bigelow, and Clara Cabezas and also our Arabic aligners, Moustafa Al-Bassyiouni, Eiman Elnahrawy, Tamer Nadeem, and Musa Nasir.

## REFERENCES

Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, I. D., Och, F.-J., Purdy, D., Smith, N. A., and Yarowsky, D. 1999. Statistical Machine Translation. Tech. rep., JHU. citeseer.nj.nec.com/al-onaizan99statistical.html.

Carbonell, J., Probst, K., Peterson, E., Monson, C., Lavie, A., Brown, R., and Levin, L. 2002. Automatic Rule Learning for Resource-Limited MT. In *Proceedings of the Fifth Conference of the Association for Machine Translation in the Americas, AMTA-2002*. Tiburon, California.

Collins, M. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. In *Proceedings of the 34th Annual Meeting of the ACL*. Santa Cruz, CA, 184–191.

Ding, Y., Gildea, D., and Palmer, M. 2003. An Algorithm for Word-Level Alignment of Parallel Dependency Trees. In *Proceedings of MT SUMMIT III*. 95–101.

Dorr, B. J., Pearl, L., Hwa, R., and Habash, N. 2002. Improved Word-Level Alignment: Injecting Knowledge about MT Divergences. Tech. rep., University of Maryland, College Park, MD. LAMP-TR-082, CS-TR-4333, UMIACS-TR-2002-15.

Gupta, D. and Chatterjee, N. 2003. Identification of Divergence for English to Hindi EBMT. In *Proceedings of MT SUMMIT III*. 141–148.

Hwa, R., Resnik, P., Weinberg, A., and Kolak, O. 2002. Evaluating Translational Correspondence using Annotation Projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA.

Lin, D. 1998. Dependency-Based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems, First International Conference on Language Resources and Evaluation*. Granada, Spain.

Lin, D. and Cherry, C. 2003. Word alignment with cohesion constraint. In *Proceedings of HLT/NAACL 2003*. Edmonton, Canada, 49–51.

Och, F. J. and Ney, H. 2000. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*. Hongkong, China, 440–447.