Rapid Porting of Divergence Unraveling (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. In this paper, we describe the adaptation of DUSTer, a divergence unraveling package, to Hindi during the Darpa TIDES-2003 Surprise Language Exercise. 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. Our ultimate 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. We present an empirical analysis comparing the complexities of performing word-level alignments with and without divergence handling. Our results suggest that our approach facilitates word-level alignment, particularly for sentence pairs containing divergences. We describe the porting process that enabled the application of DUSTer to Hindi within 3 days. We demonstrate that our approach holds promise for the rapid, large-scale acquisition of treebanked data for previously unhandled languages.

Categories and Subject Descriptors: 1.2.7 [**Artificial Intelligence**]: Natural Language Processing—machine translation

General Terms: Divergences, Machine Translation

Additional Key Words and Phrases: Divergences, Machine Translation

1. INTRODUCTION

Word-level alignments of bilingual text (bitexts) are not only an integral part of statistical machine translation models, but also useful for lexical acquisition, treebank construction, and part-of-speech tagging [Yarowsky and Ngai 2001]. The frequent occurrence of *divergences*—structural differences between languages—presents a

Authors' addresses: Bonnie J. Dorr, Department of Computer Science and UMIACS, 3153 A.V.Williams Building, University of Maryland, College Park, MD 20742.
Necip Fazil Ayan, Nizar Habash, Nitin Madnani, Institute for Advanced Computer Studies, A.V.Williams Building, University of Maryland, College Park, MD 20742.
Rebecca Hwa, Department of Computer Science, Room 5421, 210 S. Bouquet St., Pittsburgh, PA 15260.

great challenge to the alignment task. The term *divergence* refers only to differences that are relevant to predicate-argument structure as opposed to constituent re-orderings such as noun-adjective swapping which occurs between Spanish and English. See [Yamada and Knight 2001] for an approach that involves syntactic reorderings of this type.

In this paper, we describe the adaptation of DUSTer (Divergence Unraveling for Statistical Translation) to Hindi during the Darpa TIDES-2003 Surprise Language Exercise. 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.[1] Our ultimate 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. (For ease of exposition, we will henceforth refer to *non-English* as *foreign*.) The bitext is parsed on the English side only. Thus, the projected trees in the foreign language may serve as input for training parsers in a new language.

A divergence occurs when the underlying concepts or gist of a sentence are distributed over different words for different languages. For example, the notion of running into the room is expressed as *run into the room* in English and *move-in the room running* (*entrar en el cuarto corriendo*) in Spanish. While seemingly transparent for human readers, this throws statistical aligners for a serious loop. Far from being a rare occurrence, our preliminary investigations revealed that divergences occurred in approximately 1 out of every 3 sentences.[2] Thus, finding a way to deal effectively with these divergences and repair them would be a massive advance for bilingual alignment.

The following three ideas motivate the development of automatic "divergence correction" techniques:

(1) Every language pair has translation divergences that are easy to recognize.

(2) Knowing what they are and how to accommodate them provides the basis for refined word-level alignment.

(3) Refined word-level alignment results in improved projection of structural information from English to another language.

This paper elaborates primarily on points 1 and 2. Our ultimate goal is to set these in the context of 3, i.e., for training foreign-language parsers to be used in statistical machine translation.

DUSTer transforms English into a pseudo-English form (which we call E') that more closely matches the physical 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 which, in turn, facilitates our statistical alignment process. In theory, our rewriting approach applies to all

[1] See http://www.umiacs.umd.edu/labs/CLIP/DUSTer.html for more details.

[2] This analysis was done using automatic detection techniques—followed by human confirmation—on a sample size of 19K sentences from the TREC El Norte Newspaper (Spanish) Corpus, LDC catalog no LDC2000T51, ISBN 1-58563-177-9, 2000.

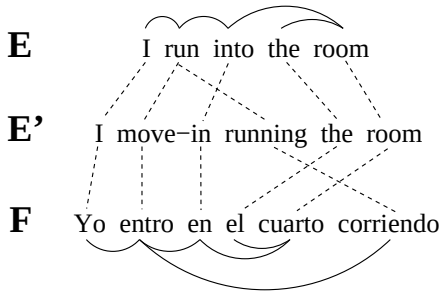

Fig. 1. Idealized Version of Transformation/Alignment/Projection

divergence types. Thus, given a corpus, divergences are identified, rewritten, and then run through the statistical aligner of choice.

The idealized version of our transformation/alignment/projection approach is illustrated for an English-Spanish pair in Figure 1. Dependencies between English words (**E**) are represented by the curves above the words—these are produced by either of two research-standard parsers, Minipar system [Lin 1995; 1998] and Collins [Collins 1996]. Alignments are indicated by dotted lines. The dependency trees are transformed into new trees associated with E', e.g., *run* and *into* in **E** are reconfigured in E' so that the sentence in E' has a one-to-one correspondence with the sentence of the foreign language **F**. The final step—outside the scope of this paper—is to induce foreign-language dependency trees automatically using statistical alignment of the E' words with those of the foreign-language sentence (e.g., using Giza++ [Al-Onaizan et al. 1999; Och and Ney 2000]).

The next section sets this work in the context of related work on alignment and projection of structural information between languages. Section 3 describes the range of divergence types covered in this work—and analyzes the frequency of their occurrence in corpora (with examples in Spanish and Arabic). Section 4 presents our methodology for porting DUSTer to a new language—in this case, Hindi—in under 3 days.[3] Section 5 describes an experiment that reveals the benefits of injecting linguistic knowledge into the alignment process. We present an empirical analysis comparing the complexities of performing word-level alignments with and without divergence handling. Our results suggest that our approach facilitates word-level alignment, particularly for sentence pairs containing divergences. We demonstrate that our approach holds promise for the rapid, large-scale acquisition of treebanked data for previously unhandled languages.

2. RELATED WORK

Recently, researchers have extended traditional statistical machine translation (MT) models [Brown et al. 1990; Brown et al. 1993] to include the syntactic structures of the languages [Alshawi and Douglas 2000; Alshawi et al. 2000; Wu 1997]. These statistical transfer systems appear to be similar in nature to what we are proposing—projecting from English to a foreign-language tree—but both the method of generation and the goal behind these approaches are different from ours. In these

[3]We found the same to be true for porting of DUSTer to Arabic and Chinese as well.

alternative approaches, parses are generated simultaneously for both sides whereas, in our approach, we assume we only have access to the English parses and then we automatically produce dependency trees in another language *without training*.[4] From these noisy foreign-language dependency trees, we then induce a parser for translation between the foreign language and English. The foreign-language parse is a necessary input to a generation-heavy decoder [Habash 2002], which produces English translations from foreign-language dependency trees.

It has been shown that MT models are significantly improved when trained on syntactically *annotated* data [Yamada and Knight 2001]. However, the cost of human labor in producing annotated treebanks is often prohibitive, thus rendering manual construction of such data for new languages infeasible. Some researchers have developed techniques for fast acquisition of hand-annotated Treebanks [Fellbaum et al. 2001]. Others have developed machine learning techniques for inducing parsers [Hermjakob and Mooney 1997; Hwa 2000], but these require extensive collections of complex translation pairs for broadscale MT.

Because divergences generally require a combination of lexical and structural manipulations, they are handled traditionally through the use of transfer rules [hye Han et al. 2000]. Unfortunately, automatic extraction of such rules relies crucially the availability of scarce resources such as large, aligned, and parsed, bilingual corpora [Lavoie et al. 2001; Menezes and Richardson 2001; Meyers et al. 2000; Watanabe et al. 2000]. Our approach requires parsing on only the English side of the aligned bilingual corpora—the foreign language need not be parsed.

Another approach that is similar to ours is that of [Carbonell et al. 2002], where MT transfer rules are learned from a set of parallel and the corresponding English parses. The difference is that our approach does not require corpus elicitation or a pre-existing alignment between the two languages.[5] More importantly, our approach imposes lexical constraints on the classes of rules that are used; these constraints are based on categories of translation divergences. For example, our rules are constrained according to parameter settings for Light Verbs and Motion Verbs.

DUSTer detects and processes divergences using linguistically motivated techniques to transform the English lexical and syntactic representation to match the physical form of the foreign language more closely—thus improving alignment. The ultimate goal is to bring about more accurate dependency-tree projection from English into the foreign language, thus producing a significantly noise-reduced dependency treebank for training foreign-language parsers.

3. FREQUENCY OF DIVERGENCES IN LARGE CORPORA

Prior to the Hindi Surprise Language exercise, we investigated divergences in Arabic and Spanish corpora to determine how often such cases arise. For Spanish, we

[4]It is important to note that rewriting the English structure as a structure in the foreign language is not intended to be an MT transfer process unto itself, but rather it is a first step in constructing a (noisy) foreign-language treebank for training a new parser for MT.

[5]However, the use of alignment for bootstrapping is currently under investigation in the DUSTer.

[6]Although cases of Head Swapping arise in Arabic and Hindi, we did not find any such cases in the small sample of sentences that we human checked.

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

Type	English	E'	Foreign Equivalent
Light Verb	fear	have fear	S: tiene miedo
	try	put to trying	S: poner a prueba
	make any cuttings	wound	A: تجرحوا
	our hand is high	our hand heightened	A: يدنا قد عظمت
	he is not here	he be-not here	A: إنه ليس هنا
	it moves at extreme speeds	it extremely be-speedy	H: वह बहुत तेज़ है
	make cuttings	wound	H: काटा
	valued at 4 rupees	value be 4 rupees	H: मूल्य चार रुपये है
Manner	teaches	walks teaching	S: anda enseñando
	is spent	self goes spending	S: se va gastando
	he sent his brothers away	he dismissed his brothers	A: صرف إخوته
	spake good of	speak-good about	A: يثني على
	he turned again	he returned	A: رجع
	the land mourns	the land stays mourning	H: धरती रोती रहती है
Structural	after six years	after of six years	S: después de seis años
	and because of that in other parts	and for that in other parts	S: y por ello en otras partes
	I forsake thee	I-forsake about-you	A: اتخلى عنك
	we found water	we-found on-water	A: عثرنا على
Categorial	I am jealous	I have jealousy	S: tengo celos
	I require of you	I require-of you	S: te pido
	(he) shall estimate	according-to (his)-estimate	A: حسب تقدير
	(how long shall) the land mourn	(stays) the-land mourning	A: الارض نائحة
	he went to	his-return to	A: عودته الى
	I am afraid	I have fear	H: मुझे डर है
	came to the beach	came PREP the beach	H: तट पर पहुँचे
	envy him	envy PREP him	H: उस से जलता है
	this is incomplete	this is not complete	H: यह पूरा नहीं है
Head-Swapping[6]	walked out	move-out walking	S: salió caminando
Thematic	I am pained	me pain they	S: me duelen
	He loves it	to him be-loved it	S: le gusta
	it was on him	he-wears-it	A: كان يرتديه
	I am pained	me pain they	H: मुझे दुख देते हैं

used TREC Spanish Data; for Arabic, we used an electronic version of the Bible written in Modern Standard Arabic. Our investigation revealed that there are six 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 six divergence types with Hindi examples. Table I shows examples of each type from our corpora, along with examples of sentence pairs. The examples shown are in A(rabic), H(indi), and S(panish).

A detailed analysis of each divergence type is given in [Dorr et al. 2002]. In a nutshell, Light Verb divergence involves the translation of a single verb to a com-

Table II. Common Search Terms for Divergence Detection

Spanish	**Arabic**
hacer (do)	ليس (be-not)
dar (give)	عبر (go-across)
tomar (take)	احسن (do-good)
tener (have)	اكتفى (make-do)
poner (put)	اخرج (take-out)
ir + X-progressive (go X-ing)	رجع (come-again)
andar + X-progressive (walk X-ing)	غرّب (go-west)
salir + X-progressive (leave X-ing)	اسرع (do-quickly)
pasar + X-progressive (pass X-ing)	جرح (make-cuttings)
entrar + X-progressive (enter X-ing)	عظم (become-high/great)
bajar + X-progressive (go-down X-ing)	صرف (send-away)
irse + X-progressive (leave X-ing)	ارتعب (be-afraid)
soler (usually)	اثني + على (speak-good + about = laud)
gustar (like)	بحث + عن (search + for = seek)
bastar (be enough)	أوصى + ب (command + with = command)
disgustar (dislike)	تخلى + عن (abandon + of = forsake)
quedar (be left over)	عثر + على (find + on = find)
doler (hurt)	عودة$_{noun}$ + الى (return$_{verb}$ to)
encantar (be enchanted by)	
importar (be important)	
interesar (interest)	
faltar (be lacking)	
molestar (be bothered by)	
fascinar (be fascinated by)	

bination of a "light" verb (carrying little or no specific meaning in its own right) and some other meaning unit (perhaps a noun) to convey the appropriate meaning. Manner divergence involves translating a single manner verb (e.g., *run*) as a light verb of motion and a manner-indicating content word. Structural divergence involves the realization of incorporated arguments such as subject and object as obliques (i.e. headed by a preposition in a PP). Categorial divergence involves a translation that uses different parts of speech. Head swapping involves the demotion of the head verb and the promotion of one of its modifiers to head position. Finally, a thematic divergence occurs when the verb's thematic roles switch syntactic arguments from one language to another.

Prior to the Hindi surprise-language exercise, we developed a set of hand-crafted regular expressions for detecting divergent sentences in our Arabic and Spanish corpora (see Table II). The Arabic regular expressions were derived by examining a small set of sentences (50), a process which took approximately 20 person-hours. The Spanish expressions were derived by a different process—involving a more general analysis of the behavior of the language—taking approximately 2 person-months. We want to emphasize that these regular expressions are *not* a sophisticated divergence detection technique. However, they do establish, at the very least, a conservative lower bound for how often divergences occur since the regular expressions pull out select cases of the different divergence types.

In this preliminary investigation, we applied the Spanish and Arabic regular expressions to a sample size of 19K Spanish sentences from TREC and 1K Ara-

Table III. Divergence Statistics

Language	Detected Divergences	Human Confirmed	Sample Size (sentences)	Corpus Size (sentences)
Spanish	11.1%	10.5%	19K	150K
Arabic	31.9%	12.4%	1K	28K

bic sentences from the Arabic Bible. Each automatically detected divergence was subsequently human verified and categorized into a particular divergence category. Table III indicates the percentage of cases we detected automatically and also the percentage of cases that were confirmed (by humans) to be actual cases of divergence.

It is important to note that these numbers reflect the techniques used to calculate them. The Arabic regular expressions were constructed more compactly than the Spanish ones in order to increase the number of verb forms that could be caught with a single expression. For example, a regular expression for a transitive verb includes the perfect and imperfect forms of the verb with various prefixes for conjugation, aspect, and tense and suffixes for pronominal direct objects. Because the Spanish regular expressions were derived through a more general language analysis, the precision is higher in Spanish than it is in Arabic. Human inspection confirmed approximately 1995 Spanish sentences out of the 2109 that were automatically detected (95% accuracy), whereas 124 sentences were confirmed in the 319 detected Arabic divergences (39% accuracy).

On the other hand, the more constrained Spanish expressions appear to give rise to a lower recall. In fact, an independent study with more relaxed regular expressions on the same 19K Spanish sentences resulted in the automatic detection of divergences in 18K sentences (95% of the corpus), 6.8K of which were confirmed by humans to be correct (35% of the corpus). Future work will involve repeated constraint adjustments on the regular expressions to determine the best balance between precision and recall for divergence detection; we believe the Arabic expressions fall somewhere in between the two sets of Spanish expressions (which are conjectured to be at the two extremes of constraint relaxation—very tight in the case above and very loose in our independent study).

The regular expressions were overgenerative in our initial investigation, i.e., our detection system encountered more seemingly divergent sentences than actually existed. Thus, we used human post-checking to eliminate erroneous cases. The remaining cases served as the basis of an analysis that allowed us to construct a "universal rule" database (described next) which we then applied directly to Hindi (and also, subsequently, to Chinese).

4. PORTING OF DUSTER TO HINDI

We accommodate Hindi divergence cases through the application of divergence transformations that are pre-stored in a repository referred to as the "universal rule" database.[7] The total number of universal rules is 117. No additional rules were added for Hindi. However, as we will see shortly, the universal rules are

[7]The universal rules are in http://www.umiacs.umd.edu/~bonnie/universal-rules.txt.

lexically *parameterized* to accommodate the new foreign language.

In the surprise-language exercise, the input to DUSTer is a set of English/Hindi sentence pairs along with the corresponding English dependency trees (produced by the Minipar system [Lin 1995; 1998] or the Collins system [Collins 1996]). DUSTer transforms the English dependency tree so that it is parallel to what would be the equivalent foreign-language dependency tree. Simultaneously, DUSTer automatically rewrites the English sentence as E'. 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 resulting E' (which was used for human alignment in the experiments to be described in Section 5) is: 'The book value(Noun) LightVB Oblique 600 rupees'. With this rewritten E' string, we can produce more accurate alignments that would otherwise be the case; these, in turn, provide the basis for more accurate projection of the English dependency tree to Hindi.

A small sample of the universal rules is shown in Table IV. The rules fall into two categories: Type I rules facilitate the task of alignment *and* enable more accurate projection of dependency trees (light verb, manner, and structural); and Type II rules *only* enable more accurate projection of dependency trees with minimal or no change to alignment accuracy (categorial, head-swapping, and thematic). In the first category, rules are sub-divided into *expansion rules*, which are applicable when the foreign language sentence is verbose relative to the English one, and *contraction rules* which are applicable when the foreign language sentence is terse relative to English.

New discoveries were made during the application of these transformation rules to Arabic, Hindi, and Spanish. In particular, we found that the expansion rules apply more frequently to Hindi and Spanish than to Arabic, whereas the reverse is true of the contraction rules. This is not surprising because, in general, Spanish is verbose relative to English, where as Arabic tends to be more terse. Such differences in verbosity are well documented in the literature. For example, according to [Slobin 1996], human translators often make changes to produce Spanish sentences that are longer than the original English sentence—or they generate sentences of the same length but *reduce* the amount of information conveyed in the original English.

As mentioned above, all E-to-E' universal rules are *parameterized* according the requirements of the left- and right-hand languages. For example, the **Light Verb** rule includes the *LightVB* parameter. The Contraction version of this rule requires access to all possible settings of this parameter in English (the left-hand language), whereas the corresponding Expansion version requires access to all possible settings of this parameter in the foreign language (the right-hand language).

The rapid setting of these parameters facilitates the porting of DUSTer to new languages: The porting process involves the use of native-speaker knowledge to set the lexical parameters encoded in the rules, e.g., LightVB, MotionV, PsychV, DirectionP, and Oblique. As an example, the setting of the LightVB parameter in English is {be, do, give, have, make, take, put}. This same parameter is set to {ser, estar, hacer, dar, tener, tomar, poner} in Spanish and { होना, करना, बनाना, लगना, लेना, डालना } in Hindi.

Table IV. Transformation Rules between E and E'

Type I. Rules Impacting Alignment and Projection

(1) **Light Verb**

Expansion: [V(PsychV) Arg1] → [V(LightVB) Arg1 N(PsychV)]
Ex: "I fear" → "I have fear"
Contraction: [V(LightVB) Arg1 Adj(DirectionV)] → [V(DirectionV) Arg1]]
Ex: "our hand is high" → "our hand heightened"

(2) **Manner**

Expansion: [V Arg1] → [V(MotionV) Arg1 V]
Ex: "I teach" → "I walk teaching"
Contraction: [V(MotionV) Arg1 Modifier] → [V(DirectionV) Arg1]
Ex: "he turns again" → "He returns"

(3) **Structural**

Expansion: [V Arg1 Arg2] → [V Arg1 P(Oblique) Arg2]
Ex: "I forsake thee" → "I forsake of thee"
Contraction: [V Arg1 P(Oblique) Arg2] → [V Arg1 Arg2]
Ex: "I search for him" → "I search him"

Type II. Rules Impacting Projection Only

(4) **Categorial**

[V Arg1 Adj(Arg2)] → [V Arg1 N(Arg2)]
Ex: "I am jealous" → "I have jealousy"

(5) **Head-Swapping**

[V(MotionV) Arg1 P(DirectionP)] → [V(DirectionV) Arg1 V(MotionV)]
Ex: "I run in" → "I enter running"

(6) **Thematic**

[V Arg1 Arg2] → [V Arg1 P(Oblique) Arg2], 1<2
Ex: "He wears it" → "It is-on him"

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

	Hindi			**Arabic**			**Chinese**		
Parameter	**Entry Count**	**Min**	**Sec/ Entry**	**Entry Count**	**Min**	**Sec/ Entry**	**Entry Count**	**Min**	**Sec/ Entry**
AspectV	14	10	43	143	15	6	581	147	15
ChangeOfStateV	408	30	4.4	1352	20	9	535	240	27
Complement	22	10	27	54	15	16	24	31	46
DirectionP	11	10	55	25	10	24	22	11	30
DirectionV	23	10	26	271	15	3	369	142	23
FunctionalDet	24	10	25	36	10	17	36	10	24
FunctionalN	46	15	20	85	15	10.5	61	24	24
LightVB	6	5	50	9	5	33	10	8	48
LocationV	75	20	16	347	15	3	181	95	30
ModalV	56	0	5.3	4	5	75	402	108	18
MotionV	87	60	41	489	15	2	96	56	35
Neg	2	1	30	8	5	37	3	2	40
Oblique	36	15	25	81	15	11	47	27	34
Pleonastic	0	0	0	6	5	50	0	0	0
PsychV	52	20	23	303	10	2	846	218	15.5
TenseV	0	0	0	10	5	30	8	4	30
Total Time	221 min = 3.7 hours = approx 0.5 days			200 min = 3.3 hours = approx 0.4 days			1029 min = 17.15 hours = approx 2.1 days		

Table V indicates the amount of time that it took to set the parameters to Hindi.[8] For comparison, we also show the time it took to develop settings for Arabic and Chinese. In each case, the parameters were set in well under 3 person-days by a native speaker of the language.

It is interesting to note that, the **more** morphologically rich languages take **less** time with respect to the setting of the associated parameters. This results in a time tradeoff that balances out all three languages in the end. We calculated the time that it took to provide morphological variants of all words stored in the parameters: 2 days for Arabic (very rich morphology), 1 day for Hindi (less rich morphology), and no time for Chinese (essentially no morphology). Thus, the overall time for incorporating a new language into DUSTer comes out to be about the same for all three languages: 2-3 days.

In addition to these parameter settings (and an English parser), DUSTer relies heavily on a large, well-developed English resource: the CatVar (categorial variation) database [Habash and Dorr 2003]. This resource serves as a trigger for certain types of part-of-speech changes (e.g., *jealous*→*jealousy* in the Categorial rule in Table IV).

Finally, DUSTer uses a set of linear ordering constraints associated with each rule, as indicated by the 1<2 specification in the thematic rule (6). (For convenience, we indicate argument ordering with "<" only in the cases where the foreign-language order does not reflect the linear order on the right-hand side of the universal rule.) These constraints are specified independently for each language.

In the surprise-language exercise, we produced E′ trees for a set of development sentence pairs from the parallel corpora provided for Hindi and English.[9] The DUSTer run took one hour to produce these results.

5. EXPERIMENT: IMPACT OF DIVERGENCE CORRECTION ON ALIGNMENT

Prior to the application of DUSTer to Hindi, we tested our hypothesis that transformations of divergent cases facilitate word-level alignment by running human alignment studies for two different pairs of languages: English-Spanish and English-Arabic. We chose these two pairings to test the generality of the divergence transformation principle. Once we ascertained the promise of the approach through

[8]The Hindi settings are available, in ITRANS format, at the following URL: http://www.umiacs.umd.edu/~bonnie/Hindi-parameters.tar.gz.

[9]The English dependency trees (produced by the Collins parser) are downloadable for use by the community:

150 English trees from BIBLE and EMILE:
http://www.umiacs.umd.edu/~bonnie/english-trees-coll-1.tar.gz

45 English trees from BIBLE and BBC:
http://www.umiacs.umd.edu/~bonnie/english-trees-coll-2.tar.gz

The original Hindi sentences for each of the English corpora above are downloadable from the following sites:
http://www.umiacs.umd.edu/~bonnie/hindi-text-1
http://www.umiacs.umd.edu/~bonnie/hindi-text-2

The corresponding E′ trees (based on Hindi) are downloadable from the following sites:
http://www.umiacs.umd.edu/~bonnie/eprime-trees-1.tar.gz
http://www.umiacs.umd.edu/~bonnie/eprime-trees-2.tar.gz

experimentation, we were well positioned to port the system over to Hindi upon the surprise-language announcement.

For each language pair, four fluently bilingual human subjects were asked to perform word-level alignments on the same set of sentences selected from the Bible. They were all provided the same instructions and software, similar to the methodology and system described by [Melamed 1998]. Two of the four subjects were given the original English and foreign language sentences; they served as the control for the experiment. The sentence given to the other two consisted of the original foreign language sentences paired with altered English (denoted as E') resulting from divergence transformations described above. We compare the inter-annotator agreement rates and other relevant statistics between the two sets of human subjects. If the divergence transformations had successfully modified English structures to match those of the foreign language, we would expect the inter-annotator agreement rate between the subjects aligning the E' set to be higher than the control set. We would also expect that the E' set would have fewer unaligned and multiply-aligned words.

Our experiment involved four steps for each foreign language L:

i. Associate the language with a subset of the "universal rules."

ii. Run DUSTer each for English sentence from a parallel set of sentences, i.e., apply the appropriate transformations to English sentence, renaming it E', according to the parametric settings for the (English, L) pair.

iii. Human alignment:

- Have two humans align the true English sentence and the foreign-language sentence.
- Have two different humans align the rewritten E' sentence and the foreign-language sentence.

iv. Compare inter-annotator agreement between the first and second sets of alignments.

5.1 Experiment 1: English and Spanish

In the case of English-Spanish, the subjects were presented with 150 sentence pairs from the English and Spanish Bibles. The sentence selection procedure is similar to the divergence detection process described in the previous section. These sentences were first selected as potential divergences, using the hand-crafted regular expressions referred to in Section 3; they were subsequently verified by the experimenter as belonging to a particular divergence type. Out of the 150 sentence pairs, 97 were verified to have contained divergences; moreover, 75 of these 97 contain expansion/contraction divergences (i.e., divergence transformations that result in altered surface words). The average length of the English sentences was 25.6 words; the average length of the Spanish sentences was 24.7 words. Of the four human subjects, two were native Spanish speakers, and two were native English speakers majoring in Spanish literature. The backgrounds of the four human subjects are summarized in Table VI.

Table VI. Summary of the backgrounds of the English-Spanish subjects

	data set	native-tongue	linguistic knowledge?	ease with computers
Subject 1	control	Spanish	yes	high
Subject 2	control	Spanish	no	low
Subject 3	divergence	English	no	high
Subject 4	divergence	English	no	low

Table VII. Summary of Backgrounds of English-Arabic Subjects

	data set	native-tongue	linguistic knowledge?	ease with computers
Subject 1	control	Arabic	yes	high
Subject 2	control	Arabic	no	high
Subject 3	divergence	Arabic	no	high
Subject 4	divergence	Arabic	no	high

Table VIII. Results of Two Experiments on All Sentence Pairs[10]

	# of sentences	F-score	% of unaligned words	Avg. alignments per word
E-S	150	80.2	17.2	1.35
E'-S	150	82.9	14.0	1.16
E-A	50	69.7	38.5	1.48
E'-A	50	75.1	11.9	1.72

Table IX. Results for Subset Containing only Divergent Sentences

	# of sentences	F-score	% of unaligned words	Avg. alignments per word
E-S	97	81.0	17.3	1.35
E'-S	97	83.8	13.8	1.16
E-A	50	69.7	38.5	1.48
E'-A	50	75.1	11.9	1.72

Table X. Results for Subset of Sentence Pairs with only Expansion/Contraction Divergences

	# of sentences	F-score	% of unaligned words	Avg. alignments per word
E-S	75	82.2	17.3	1.34
E'-S	75	84.6	13.9	1.14
E-A	36	69.1	38.3	1.48
E'-A	36	75.7	11.5	1.67

5.2 Experiment 2: English and Arabic

In the case of English-Arabic, the subjects were presented with 50 sentence pairs from the English and Arabic Bibles. While the total number of sentences was smaller than the previous experiment, every sentence pair was verified to contain at least one divergence. Of these 50 divergent sentence pairs, 36 of them contained expansion/contraction divergences. The average English sentence length was 30.5 words, and the average Arabic sentence length was 17.4 words. The backgrounds of the four human subjects are summarized in Table VII.

Inter-annotator agreement rate is quantified for each pair of subjects who viewed the same set of data. We hold one subject's alignments as the "ideal" and compute the precision and recall figures for the other subject based on how many alignment links were made by both people. The averaged precision and recall figures (F-scores)[11] for the the two experiments and other relevant statistics are summarized in Table VIII. In both experiments, the inter-annotator agreement is higher for the bitext in which the divergent portions of the English sentences have been transformed. For the English-Spanish experiment, the agreement rate increased from 80.2% to 82.9% (error reduction of 13.6%). Using the pair-wise t-test, we find that the higher agreement rate is statistically significant with 95% confidence. For the English-Arabic experiment, the agreement rate increased from 69.7% to 75.1% (error reduction of 17.8%); this higher agreement rate is statistically significant with a confidence rate of 90%.

We also performed data analyses on two subsets of the full study. First, we focused on sentence pairs that were verified to contain divergences; the results are reported in Table IX. They were not significantly different from the complete set. We then considered a smaller subset of sentence pairs containing only expansion/contraction divergences whose transformations altered the surface words as well as the syntactic structures; the results are reported in Table X. In this case, the higher agreement-rate for the E'-Spanish annotators is statistically significant with 90% confidence; the higher agreement-rate for the E'-Arabic annotators is statistically significant with 95% confidence.

Additional statistics also support our hypothesis that transforming divergent English sentences facilitates word-level alignment by reducing the number of unaligned and multiply-aligned words. In the English-Spanish experiment, both the appearances of unaligned words and multiply-aligned words decreased when aligning to the modified English sentences. The percentage of unaligned words decreased from 17% to 14% (18% fewer unaligned words), and the average number of links to a word is lowered from 1.35 to 1.16.[12] In the English-Arabic experiment, the number of unaligned words is significantly smaller when aligning Arabic sentences to the modified English sentences; however, on average multiple-alignment increased. This may be due to the big difference in sentence lengths (English sentences are typically twice as long as the Arabic ones); thus it is not surprising that the average number of alignments per word would be closer to two when most of the words are aligned. The reason for the lower number in the unmodified English case might be that the subjects only aligned words that had clear translations.

6. CONCLUSION AND FUTURE WORK

We have shown that divergence cases can be systematically handled by transforming the syntactic structures of the English sentences to bear a closer resemblance to those of the foreign language, using a small set of templates. The validity of the divergence handling has been verified through two word-level alignment exper-

[10] In computing the average number of alignments per word, we do not include unaligned words.

[11] $F = \frac{2 \times Precision \times Recall}{Precision + Recall}$

[12] The relatively high overall percentage of unaligned words is due to the fact that the subjects did not align punctuation marks.

iments. In both cases, the human subjects consistently had higher agreement rate with each other on the task of performing word-level alignment when divergent English phrases were transformed.

The results of this work suggest several future research directions. Currently we are running human alignment experiments for Hindi and English to determine whether we achieve the same degree of inter-annotator agreement among our human subjects that we achieved with our Spanish and Arabic experiments. A pilot is underway for this experiment, using the 195 Hindi-English sentence pairs mentioned in Section 4. As before, half of our human subjects will align English sentences to the original Hindi sentences; the other half will align E′ sentences to the original Hindi sentences. (The results for this experiment will be obtained within 2 weeks.) We will then determine the degree to which DUSTer improves automatic alignment, using a much larger test set provided on the Hindi surprise-language website. We will apply Giza++ [Al-Onaizan et al. 1999; Och and Ney 2000] to English and Hindi and also to the corresponding E′ and Hindi. We will then evaluate whether the English/Hindi alignments induced from the automatic E′/Hindi alignments are more accurate than the English/Hindi alignments produced directly by Giza++. 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.

While we have focused on the effect of divergence handling on the word-alignment process in this work, we also need to evaluate the effect of divergence handling on the foreign parse trees. Our latest experiments involve projection of English trees to Chinese; we will evaluate whether our transformation rules on the English structures result in better projected Chinese dependency structures by evaluating against Chinese Treebank data [Xia et al. 2000].

Finally, we plan to compare our approach with that of [Hwa et al. 2002] in creating foreign language treebanks from projected English syntactic structures. Both approaches apply techniques to improve the accuracy of projected dependency trees, but ours occur *prior* to statistical alignment, making corrections relevant to general classes of divergences—whereas the latter occurs *after* statistical alignment, making corrections relevant to syntactic constraints of the foreign language. We will evaluate different orderings of the two different correction types to determine which ordering is most appropriate for optimal projection of foreign-language dependency trees.

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.

ALSHAWI, H., BANGALORE, S., AND DOUGLAS, S. 2000. Learning Dependency Translation Models as Collections of Finite State Head Transducers. *Computational Linguistics 26*.

ALSHAWI, H. AND DOUGLAS, S. 2000. Learning Dependency Transduction Models from Unannotated Examples. *Philosophical Transactions, Series A: Mathematical, Physical and Engineering Sciences*.

BROWN, P. F., COCKE, J., DELLA-PIETRA, S., DELLA-PIETRA, V. J., JELINEK, F., LAFFERTY, J. D., MERCER, R. L., AND ROOSSIN, P. S. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics 16,* 2 (June), 79–85.

BROWN, P. F., DELLAPIETRA, S. A., DELLAPIETRA, V. J., AND MERCER, R. L. 1993. The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*.

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.

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.

FELLBAUM, C., PALMER, M., DANG, H. T., DELFS, L., AND WOLFF, S. 2001. Manual and Automatic Semantic Annotation with WordNet. In *Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources: Applications, Customizations*. Carnegie Mellon University, Pittsburg, PA.

HABASH, N. 2002. Generation-Heavy Machine Translation. In *Proceedings of the International Natural Language Generation Conference (INLG'02) Student Session*. New York.

HABASH, N. AND DORR, B. J. 2003. A Categorial Variation Database for English. In *Proceedings of North American Association for Computational Linguistics*. Edmonton, Canada.

HERMJAKOB, U. AND MOONEY, R. J. 1997. Learning Parse and Translation Decisions from Examples with Rich Context. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, 482–489.

HWA, R. 2000. Sample selection for statistical grammar induction. In *Proceedings of the 2000 Joint SIGDAT Conference on EMNLP and VLC*. Hong Kong, China, 45–52.

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.

HYE HAN, C., LAVOIE, B., PALMER, M., RAMBOW, O., KITTREDGE, R., KORELSKY, T., KIM, N., AND KIM, M. 2000. Handling Structural Divergences and Recovering Dropped Arguments in a Korean/English Machine Translation System. In *Proceedings of the Fourth Conference of the Association for Machine Translation in the Americas, AMTA-2000*. Cuernavaca, Mexico.

LAVOIE, B., WHITE, M., AND KORELSKY, T. 2001. Inducing Lexico-Structural Transfer Rules from Parsed Bi-texts. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics – DDMT Workshop*. Toulouse, France.

LIN, D. 1995. Government-Binding Theory and Principle-Based Parsing. Tech. rep., University of Maryland. Submitted to Computational Linguistics.

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.

MELAMED, I. D. 1998. Empirical Methods for MT Lexicon Development. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*. Springer-Verlag, Langhorne, PA, 18–30. `ftp://ftp.cis.upenn.edu/pub/melamed/papers/amta98.ps.gz`.

MENEZES, A. AND RICHARDSON, S. D. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics – DDMT Workshop*. Toulouse, France.

MEYERS, A., KOSAKA, M., AND GRISHMAN, R. 2000. Chart-Based Transfer Rule Application in Machine Translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000)*. Saarbrucken, Germany.

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.

SLOBIN, D. I. 1996. Two Ways to Travel: Verbs of Motion in English and Spanish. In *Grammatical Constructions: Their Form and Meaning*, M. Shibatani and S. A. Thompson, Eds. Oxford University Press, New York, 195–219.

WATANABE, H., KUROHASHI, S., AND ARAMAKI, E. 2000. Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Transaltion. In *Proceedings of COLING-2000*. Saarbrüken, Germany.

WU, D. 1997. Stochastic Inversion Transduction Grammars and Bilingual Parsing of Parallel Corpora. *Computational Linguistics 23,* 3, 377–400.

XIA, F., PALMER, M., XUE, N., OCUROWSKI, M. E., KOVARIK, J., CHIOU, F.-D., HUANG, S., KROCH, T., AND MARCUS, M. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. In *Proceedings of the Second Language Resources and Evaluation Conference*. Athens, Greece.

YAMADA, K. AND KNIGHT, K. 2001. A Syntax-Based Statistical Translation Model. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, 523–529.

YAROWSKY, D. AND NGAI, G. 2001. Inducing Multilingual POS Taggers and NP Bracketers via Robust Projection across Aligned Corpora. In *Proc. of NAACL-2001*. 200–207.