Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation

Nizar Habash and Bonnie Dorr

Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20740
{habash,bonnie}@umiacs.umd.edu
http://umiacs.umd.edu/labs/CLIP
Track: Theoretical Paper

Abstract. This paper describes a novel approach to handling translation divergences in a Generation-Heavy Hybrid Machine Translation (GHMT) system. The translation divergence problem is usually reserved for Transfer and Interlingual MT because it requires a large combination of complex lexical and structural mappings. A major requirement of these approaches is the accessibility of large amounts of explicit symmetrical knowledge for both source and target languages. This limitation renders Transfer and Interlingual approaches ineffective in the face of structurally-divergent language pairs with asymmetrical resources. GHMT addresses the more common form of this problem, source-poor/target-rich, by fully exploiting symbolic and statistical target-language resources. This is accomplished by using target-language lexical semantics, categorial variations and subcategorization frames to overgenerate multiple lexico-structural variations from a target-glossed syntactic dependency of the source-language sentence. The symbolic overgeneration, which accounts for different possible translation divergences, is constrained by a statistical target-language model.

1 Introduction

In this paper, we describe a novel approach to handling translation divergences using the Generation-Heavy Hybrid Machine Translation (GHMT) model introduced in [7]. The translation divergence problem is usually reserved for Transfer and Interlingual MT because it requires a large combination of complex lexical and structural mappings. A major requirement of these approaches is the accessibility of large amounts of explicit symmetrical knowledge for both the source language (SL) and the target language (TL). This limitation makes Transfer and Interlingua inapplicable approaches to structurally-divergent language pairs with asymmetrical resources. GHMT addresses the more common form of this problem, source-poor/target-rich, by fully exploiting symbolic and statistical TL resources.

SLs are only expected to have a syntactic parser and a translation lexicon that maps SL words to TL bags of words. No transfer rules or complex interlingual representations are required. The approach depends on the existence of rich TL resources such as lexical semantics, categorial variations and subcategorization frames to overgenerate multiple lexico-structural variations from a target-glossed syntactic dependency of the SL sentence. The symbolic overgeneration, which accounts for different possible translation divergences, is constrained by a statistical TL model.

The work presented here focuses on the generation component of GHMT and its handling of translation divergences. The next section describes the range of divergence types covered in this work and discusses previous approaches to handling them in MT. Section 3 describes the components of the GHMT approach. Finally, Section 4 addresses the interaction between statistical and symbolic knowledge in the system through illustrative examples.

2   Background: Translation Divergences

A translation divergence occurs when the underlying concept or "gist" of a sentence is distributed over different words for different languages. For example, the notion of floating across a river is expressed as *float across a river* in English and *cross a river floating* (*atravesó el río flotando*) in Spanish [3]. An investigation done by [4] found that divergences occurred in approximately 1 out of every 3 sentences in the TREC El Norte Newspaper Corpus[1]. In the next section, we describe translation divergence types before turning to alternative approaches to handling them.

2.1   Translation Divergence Types

While there are many ways to classify divergences, we present them here in terms of five specific divergence *types* that can take place alone or in combination with other types of translation divergences. Table 1 presents these divergence archetypes with Spanish-English examples.[2]

- *Categorial Divergence*: Categorial divergence involves a translation that uses different parts of speech.
- *Conflation*: Conflation involves the translation of two words using a single word that combines their meaning. In Spanish-English translation, this divergence type usually involves a single English verb being translated using a combination of a light verb[3], and some other meaning-heavy unit such as a noun or a progressive manner verb.

---

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

[2] The divergence categories are described in more detail in [4].

[3] Semantically "light" verbs carry little or no specific meaning in their own right such as *give*, *do* or *have*.

Divergence	Spanish		English	%
Categorial	X *tener hambre*	(X *have hunger*)	X *be hungry*	98%
	X *tener celos*	(X *have jealousy*)	X *be jealous*	
Conflational	X *dar puñaladas a* Z	(X *give stabs to* Z)	X *stab* Z	83%
	X *ir pasando*	(X *go passing*)	X *pass*	
Structural	X *entrar en* Y	(X *enter in* Y)	X *enter* Y	35%
	X *pedir* un referendum	(X *ask-for* a referendum)	X *ask for* a referendum	
Head Swapping	X *cruzar* Y *nadando*	(X *cross* Y *swimming*)	X *swim across* Y	8%
	X *entrar corriendo*	(X *enter running*)	X *run in*	
Thematic	X *gustar a* Y	(X *please to* Y)	Y *like* X	6%
	X *doler a* Y	(X *hurt to* Y)	Y *hurt* from X	

Table 1. Translation Divergence Types

- *Structural Divergence*: A structural divergence involves the realization of incorporated arguments such as subject and object as obliques (i.e. headed by a preposition in a PP) or vice versa.
- *Head Swapping*: This divergence involves the demotion of the head verb and the promotion of one of its modifiers to head position. In other words, a permutation of semantically equivalent words is necessary to go from one language to the other. In Spanish, this divergence is typical in the translation of an English motion verb and a preposition as a directed motion verb and a progressive verb.
- *Thematic Divergence*: A thematic divergence occurs when the verb's arguments switch syntactic argument roles from one language to another (i.e. subject becomes object and object becomes subject). The Spanish verbs *gustar* and *doler* are examples of this case.

The last column in Table 1 displays a percentage of occurrences of the specific divergence type, taken from the first 48 unique instances of Spanish-English divergences from the TREC El Norte corpus. Note that there is often overlap among the divergence types with the categorial divergence occurring almost every time there is any other type of divergence. An extreme example of divergence type cooccurrence is *Maria tiene gustos de políticos diferentes*, which can be translated as *different politicians please Maria*. There four are divergence types in this pair: categorial ($gusto_{noun}$ to $please_{verb}$), conflational (*tener gusto* to *please*), thematic (*Maria* and *politicians* switch syntactic roles) and structural (*politican* is an oblique in Spanish but an argument in English). This highlights the need for a systematic approach to handling divergences that addresses all their different types and the interactions amongst them rather than addressing specific cases one at a time.

2.2   Handling Translation Divergences

Since translation divergences require a combination of lexical and structural manipulations, they are traditionally handled minimally through the use of transfer

rules [8, 15]. A pure transfer approach is a brute force attempt to manually encode all translation divergences in a transfer lexicon [5]. Very large parsed and aligned bilingual corpora have also been used to automatically extract transfer rules [16, 20]. This approach depends on the availability of such resources, which are very scarce. Alternatively, more linguistically-sophisticated techniques that use lexical semantic knowledge to detect and handle divergences have been developed.

One approach uses Jackendoff's Lexical Semantic Structure (LCS) [9, 10] as an interlingua [3]. LCS is a compositional abstraction with language-independent properties that transcend structural idiosyncrasies by providing a granularity of representation much finer than syntactic representation. LCS has been used in several projects such as UNITRAN [2] and ChinMT [19]. As an example, the Spanish sentence *Juan cruza el río nadando* can be "composed" as the following LCS using a Spanish LCS lexicon as part of the interlingual analysis:

(1) `[event CAUSE JOHN`
      `[event GO JOHN [path ACROSS JOHN [position AT JOHN RIVER]]]`
      `[manner SWIM+INGLY]]`

In the generation phase, this same LCS is "decomposed" using English LCS lexicon entries to yield *John swam across the river*.

Another approach enriches lexico-structural transfer at Mel'čuk's Deep Syntactic Structure (DSyntS) level [17] with cross-linguistic lexical semantic features [18]. Transfer lexicon rules are written as such to capture generalizations across the language pair instead of addressing specific paired instances. As an example, the following transfer rule can be used to handle the head swapping divergence discussed in the last example.

(2) `@TRANS_CORR`
    `@EN V1 [cat:verb manner:M]`
      `(ATTR Y [cat:prep path:P event:go] (II N))`
    `@SP V2 [cat:verb path:P event:go]`
      `(II N ATTR Z [manner:M])`

Here, a transfer correspondence is established between the different components of two DSyntS templates. Note how the manner variable M and the path variable P switch dominance.

A major limitation of these interlingual and transfer approaches (whether using lexical semantics or corpus-based) is that they require a large amount of explicit symmetrical knowledge for both SL and TL. The Generation-Heavy Machine Translation approach (GHMT) is closely related to the hybrid approach described in [11, 12, 13]. The idea is to combine symbolic and statistical knowledge in generation through a two step process: (1) Symbolic Overgeneration followed by (2) Statistical Extraction. The hybrid approach has been for generation from semantic representations [12] or from shallow unlabeled dependencies [1]. GHMT extends on earlier work by including structural and categorial expansion of SL syntactic dependencies as part of the symbolic overgeneration component.
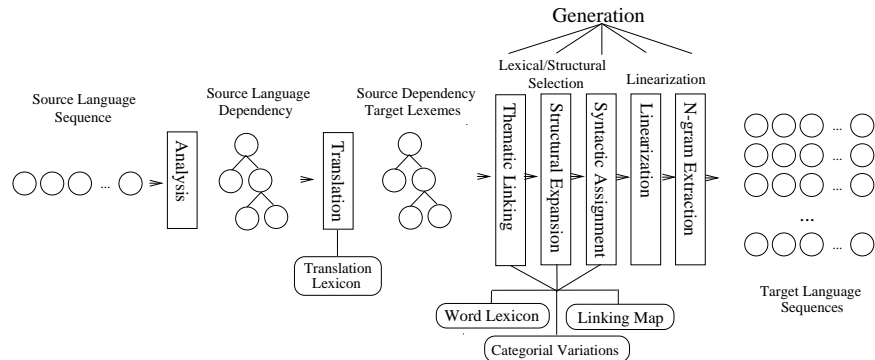
Fig. 1. Generation-Heavy Machine Translation

The fact that GHMT doesn't require semantically analyzed SL representations or structural transfer lexicons, makes it perfect for handling translation divergences with relatively minimal lexical resources on for the SL. The overgeneration is constrained by linguistically-motivated rules that utilize TL lexical semantics and is independent of the SL preferences. The generated lexico-structural combinations are then ranked by the statistical extraction component. Figure 1 presents an overview of the complete MT system.

3 Generation-Heavy Machine Translation

The three phases of GHMT—Analysis, Translation and Generation—are very similar to other paradigms of MT: Analysis-Transfer-Generation or Analysis-Interlingua-Generation [5]. However, these phases are not symmetrical. Analysis relies only on the SL sentence parsing and is independent of the TL. The output of Analysis is a deep syntactic dependency that normalizes over syntactic phenomena such as passivization and morphological expressions of tense, number, etc. Translation converts the SL lexemes into bags of TL lexemes. The dependency structure of the SL is maintained. The last phase, Generation, is where most of the work is done to manipulate the input lexically and structurally and produce TL sequences. Next we will describe the generation resources followed by an explanation of the generation sub-modules.

3.1 Generation Resources

The generation component utilizes three major TL resources (see Figure 1). First, the word-class lexicon defines verbs and prepositions in terms of their subcategorization frames and lexical conceptual primitives. A single verb or preposition can have multiple entries for each of its senses. For example, among other entries, run_1 as in ($John_{agent}$ $ran_{cause-go}$ $store_{theme}$) is distinguished from run_2 as in ($John_{theme}$ ran_{go}). Second, the categorial-variation lexicon relates words

to their categorial variants. For example, $hunger_V$, $hunger_N$ and $hungry_{AJ}$ are clustered together. So are $cross_V$ and $across_P$; and $stab_V$ and $stab_N$. Finally, the syntactic-thematic linking map relates syntactic relations (such as subject and object) and prepositions to the thematic roles they can assign. For example, while a subject can take on just about any thematic role, an indirect object is typically a *goal, source* or *benefactor*. Prepositions can be more specific. For example, *toward* typically marks a *location* or a *goal*, but never a *source*.

3.2   Generation Sub-modules

The generation component contains five steps (Figure 1). The first three are responsible for lexical and structural selection and the last two are responsible for linearization. We focus on these first three steps here, but see [7] for a description of the last two steps.

Initially, the SL syntactic dependency—now containing TL lexemes—is converted into a thematic dependency. The syntax-thematic linking is achieved through the use of thematic grids associated with English (verbal) head nodes together with the syntactic-thematic linking map. This step is a *loose* linking step that does not enforce the subcategorization-frame ordering or preposition specification. This looseness is important for linking from non-English subcategorization frames. For example, although the sentence **Mary filled water in the glass* is a bad English sentence (albeit good Korean), its arguments are mapped correctly as *agent, theme* and *location* respectively. The correct mapping is reached because *in* is a *location*-specifying preposition.

The next step is structural expansion, which explores conflated and head-swapped variations of the thematic dependency. Conflation is handled by examining all verb-argument pairs (V_{head}, Arg) for *conflatability*. For example, in *John put salt on the butter, to put salt on* can be conflated as *to salt* but *to put on butter* cannot be conflated into *to butter*. The thematic relation between the argument and its head together with other lexical semantic features constrain this structural expansion. Head Swapping is restricted through a similar process that examines head-modifier pairs for *swappability*.

The third step turns the thematic dependency into a full TL syntactic dependency. Syntactic positions are assigned to thematic roles using the verb class subcategorization frames and argument category specifications.

These three steps in the generation component address different translation divergence types. The thematic linking normalizes the input with respect to the thematic and structural divergences. Once the thematic roles are identified and surface syntactic cases are invisible, structural expansion can take place to handle conflation and head-swapping possibilities. The very common categorial divergence is handled at the structural expansion step too, but it is also fully addressed in the syntactic assignment step.

Finally, in the linearization step, a rule based grammar is used to create a word lattice that encodes the different possible realizations of the sentence. The grammar is implemented using the linearization engine oxyGen [6]. Sentences

6

are ranked with Nitrogen's Statistical Extractor using a uni/bigram model of two years of Wall Street Journal [13].

4 Discussion: Symbolic-Statistical Knowledge and Translation Divergences

A preliminary evaluation of GHMT conducted by [7] found that four of every five Spanish-English divergences can be generated using structural expansion and categorial variations[4]. Here we look at the interaction between symbolic and statistical knowledge[5] in GHMT in the context of divergence handling using the following two illustrative Spanish-English divergent examples:

(3) Juan tiene hambre.

```
John is hungry .    [ LENGTH 4, SCORE -7.111878 ]
John hunger .   [ LENGTH 3, SCORE -7.474780 ]
John is starved .   [ LENGTH 4, SCORE -7.786015 ]
John is a hunger .    [ LENGTH 5, SCORE -8.173432 ]
John has a hunger .   [ LENGTH 5, SCORE -8.613148 ]
John is a famine .    [ LENGTH 5, SCORE -8.666368 ]
John is the hunger .    [ LENGTH 5, SCORE -8.829170 ]
John is the famine .    [ LENGTH 5, SCORE -8.871368 ]
John be hungry .    [ LENGTH 4, SCORE -9.038840 ]
John is a starvation .    [ LENGTH 5, SCORE -9.105497 ]
```

(4) Yo le di puñaladas a Juan.

```
I stabbed John .    [ LENGTH 4, SCORE 0.670270 ]
I gave a stab at John .    [ LENGTH 7, SCORE -2.175831 ]
I gave the stab at John .    [ LENGTH 7, SCORE -3.969686 ]
I gave an stab at John .    [ LENGTH 7, SCORE -4.489933 ]
I gave a stab by John .    [ LENGTH 7, SCORE -4.803054 ]
I gave a stab to John .    [ LENGTH 7, SCORE -5.045810 ]
I gave a stab into John .    [ LENGTH 7, SCORE -5.810673 ]
I gave a stab through John .    [ LENGTH 7, SCORE -5.836419 ]
I gave a knife wound by John .    [ LENGTH 8, SCORE -6.041891 ]
I gave John a knife wound .    [ LENGTH 7, SCORE -6.212851 ]
```

The correct form of the output is ranked highest in both cases: *John is hungry* and *I stabbed John*. However, the ranking doesn't reflect fluency or accuracy. For example, *I gave John a knife wound* ranks much lower than *I gave an stab at John* although the former is more fluent. And the generation of *John is a hunger* as a variant of *John is hungry* is an inaccurate translation. These problems can be

---

[4] The rest of the cases require more conceptual knowledge, pragmatic knowledge and/or hard-wiring of idiomatic non-decompositional expressions.

[5] For a general discussion of the value of statistical knowledge in hybrid systems, see [14].

blamed either on the symbolic component's overgeneration or on the statistical component's *under*-extraction.

One case highlighting the issue of fluency is the generation of the sequence *John hunger* in example (3). Here, the symbolic rules are not enforcing any subject-verb agreement, which results in allowing the sequence *John hunger* in the generated word lattice together with *John hungers*. However, the statistical model fails in at least ranking *John hunger* lower than *John hungers*, which doesn't even make it to the top-ten sequences. This failure is likely due to the smoothing model used for handling unseen bigrams, which depends on the word unigrams instead. *Hunger* is a more common unigram than *hungers*.

Another case relevant to the fluency issue is the underspecification of preposition selection for the verb *give* in example (4). The current constraint is that the selected preposition *could* assign the thematic role, in this case, *goal*. Thus, the preposition *by* selected for *I gave a stab by John* has the locational not the agentive sense. The statistical model failure here is likely due to uni/bigrams enforcing fluency locally on a very small window. A possible solution on the statistical side is to use structural n-grams that capture long-distance dependencies between the verb and its modifiers.

The case of generating *John is a hunger* in example (3) reflects the dependency of GHMT on TL statistical knowledge as opposed to translingual knowledge of translation divergences. The argument here is that generating the metaphoric *John is a hunger* is a "compromise" of accuracy worth taking when generated with more likely sequences such as *John is hungry*. If the SL input was a metaphoric *John BE hunger*, then other verbs besides *be* would not be generated to start with and the smaller search space will allow the less likely metaphoric expression to be selected. This argument is, of course, hard to evaluate—and in example (3), *John is a hunger* ranks higher than the poetic *John has a hunger*. This ordering is a result of the statistical extraction use of bigrams in our current system which picks *John is a X* over *John has a X* regardless of *X*.

5   Conclusions and Future Work

There is a wide room for improving the system's correctness. Stricter symbolic rules can be implemented to limit extraneous overgeneration. A better statistical language model, one that models dependency relations (similar to [1]), can be used to rule out additional cases by enforcing long-distance dependency relations. The interaction between the two components is open for further research. An important point to note here is that all proposed modifications to these two components make use of *only* TL knowledge and resources, which guarantees the SL independence claim of GHMT.

Our immediate future work will involve an expansion of the linearization grammar to handle large-scale Spanish-English GHMT. We will conduct a more extensive evaluation of the behavior of the system as a whole including a comparative analysis of other models of Spanish-English MT (an interlingual model

and a transfer model). We also plan to explore extensions to the statistical component through the use of structural bigrams. A more extensive evaluation of the behavior of the system as a whole including a comparative analysis of other models of MT that handle divergences is planned as well. And finally, we are interested in testing our SL independence claim by retargeting the system to Chinese input.

Acknowledgments

References

[1] S. Bangalore and O. Rambow. Exploiting a Probabilistic Hierarchical Model for Generation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrcken, Germany, 2000.

[2] Bonnie J. Dorr. Interlingual Machine Translation: A Parameterized Approach. *Artificial Intelligence*, 63(1 & 2):429–492, 1993.

[3] Bonnie J. Dorr. *Machine Translation: A View from the Lexicon*. The MIT Press, Cambridge, MA, 1993.

[4] Bonnie J. Dorr. Improved Word-Level Alignment: Injecting Knowledge about MT Divergences. Technical report, University of Maryland, College Park, MD, 2002. LAMP-TR-082, CS-TR-4333, UMIACS-TR-2002-15.

[5] Bonnie J. Dorr, Pamela W. Jordan, and John W. Benoit. A Survey of Current Research in Machine Translation. In M. Zelkowitz, editor, *Advances in Computers, Vol. 49*, pages 1–68. Academic Press, London, 1999.

[6] Nizar Habash. oxyGen: A Language Independent Linearization Engine. In *Fourth Conference of the Association for Machine Translation in the Americas, AMTA-2000*, Cuernavaca, Mexico, 2000.

[7] Nizar Habash. Generation-Heavy Hybrid Machine Translation. In *Proceedings of the International Natural Language Generation Conference (INLG'02)*, New York, 2002.

[8] Chung hye Han, Benoit Lavoie, Martha Palmer, Owen Rambow, Richard Kittredge, Tanya Korelsky, Nari Kim, and Myunghee Kim. 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, 2000.

[9] Ray Jackendoff. *Semantics and Cognition*. The MIT Press, Cambridge, MA, 1983.

[10] Ray Jackendoff. *Semantic Structures*. The MIT Press, Cambridge, MA, 1990.

[11] K. Knight and V. Hatzivassiloglou. Two-Level, Many-Paths Generation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, pages 252–260, Cambridge, MA, 1995.

[12] Irene Langkilde and Kevin Knight. Generating Word Lattices from Abstract Meaning Representation. Technical report, Information Science Institute, University of Southern California, 1998.

[13] Irene Langkilde and Kevin Knight. Generation that Exploits Corpus-Based Statistical Knowledge. In *ACL/COLING 98, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (joint with the 17th International Conference on Computational Linguistics)*, pages 704–710, Montreal, Canada, 1998.

[14] Irene Langkilde and Kevin Knight. The Practical Value of N-Grams in Generation. In *International Natural Language Generation Workshop*, 1998.

[15] Benoit Lavoie, Richard Kittredge, Tanya Korelsky, and Owen Rambow. A Framework for MT and Multilingual NLG Systems Based on Uniform Lexico-Structural Processing. In *Proceedings of the 1st Annual North American Association of Computational Linguistics, ANLP/NAACL-2000*, Seattle, WA, 2000.

[16] Benoit Lavoie, Michael White, and Tanya Korelsky. 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, 2001.

[17] Igor Mel'čuk. *Dependency Syntax: Theory and Practice*. State University of New York Press, New York, 1988.

[18] Alexis Nasr, Owen Rambow, Martha Palmer, and Joseph Rosenzweig. Enriching Lexical Transfer With Cross-Linguistic Semantic Features (or How to Do Interlingua without Interlingua). In *Proceedings of the 2nd International Workshop on Interlingua*, San Diego, California, 1997.

[19] David Traum and Nizar Habash. Generation from Lexical Conceptual Structures. In *Proceedings of the Workshop on Applied Interlinguas, North American Association of Computational Linguistics/Applied Natural Language Processing Conference, NAACL/ANLP-2000*, pages 34–41, Seattle, WA, 2000.

[20] Hideo Watanabe, Sadao Kurohashi, and Eiji Aramaki. Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Transaltion. In *Proceedings of COLING-2000*, Saarbrüken, Germany, 2000.