Generation-Heavy Hybrid Machine Translation

Nizar Habash
Institute of Advanced Computer Studies
University of Maryland
College Park, MD 20740
habash@umiacs.umd.edu

Abstract

This paper describes Generation-Heavy Hybrid Machine Translation (GHMT), a novel approach for translating between structurally-divergent language pairs with asymmetrical resources. The approach depends on the existence of rich target language resources such as word lexical semantics, categorial variations and subcategorization frames. These resources are used to overgenerate multiple lexico-structural variations from a target-glossed syntactic dependency representation of the source language sentence. This symbolic overgeneration, which accounts for a wide range of possible variations, is constrained by a statistical target-language model. The exploitation of target language resources (symbolic and statistical) to handle a problem usually reserved for Transfer and Interlingual MT is useful for translation from source languages with scarce linguistic resources. A preliminary evaluation on the application of this approach to Spanish-English MT is conducted with promising results.

1 Introduction

Generation-Heavy Machine Translation (GHMT) is a novel approach for translating between structurally-divergent language pairs with asymmetrical resources. In this model, the generation component is what constrains the translation using a combination of symbolic rules, lexicons, and corpus-based statistics. The source language (SL) is only expected to have a syntactic parser and a simple one-to-many translation lexicon. No transfer rules or complex interlingual representations are used. The approach depends on the existence of rich target language (TL) resources such as word lexical semantics, categorial variations and subcategorization frames. These resources are used to generate multiple structural variations from a target-glossed syntactic dependency representation of SL sentences. This symbolic overgeneration, which accounts for possible translation divergences, is constrained by a statistical TL model. The exploitation of TL resources (symbolic and statistical) to handle a problem usually reserved for Transfer and Interlingual MT is useful for translation from structurally divergent SLs with scarce linguistic resources. A preliminary evaluation on the application of this approach to Spanish-English MT proves it extremely promising.

The next section describes the range of divergence types covered in this work and discusses previous approaches to handling them. Section (3) describes the different components and algorithms in the translation system. And finally, Section (4) describes a preliminary evaluation undertook to assess the applicability of this approach to Spanish-English MT.

2 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 (Dorr, 1993). An investigation done by (Dorr et al., 2002) found that

divergences occurred in approximately 1 out of every 3 sentences in a sample size of 19K sentences from the TREC El Norte Newspaper Corpus. This analysis was done on the TREC Spanish Data[1] using automatic detection techniques followed by human confirmation.

2.1 Translation Divergence Types

Translation divergences can be classified in terms of five specific divergence *types* that can take place alone or jointly.

1. The categorial divergence involves a translation that uses different parts of speech, e.g., 'hungry' as 'hunger'. This is by far the most common divergence type, overlapping almost completely with all other divergence types.

2. The conflational divergence involves the translation of two words using a single word that combines their meaning, e.g., 'stab' as 'give stabs' or 'butter' as 'put butter'.

3. The 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.

4. The head swapping divergence involves the demotion of the head verb and the promotion of one of its modifiers to head position. In Spanish, this divergence is typical in the translation of an English motion verb (e.g. 'float') and a preposition (e.g. 'across') as a directed motion verb and a progressive verb ('cross floating').

5. The thematic divergence occurs when the linking between syntactic arguments and thematic roles is switched during the translation from one language to another. The Spanish verbs *gustar* ('to like') and *doler* ('to hurt') are examples of this case.

2.2 Handling Translation Divergences

Since translation divergences require a combination of lexical and structural manipulation, they are traditionally handled at the transfer or interlingual levels of the MT Hierarchy. A pure brute-force transfer approach attempts to encode all translation divergences in a lexicon of transfer rules (Han et al., 2000). Very large parsed and aligned bilingual corpora have also been used to automatically extract transfer rules (Lavoie et al., 2001). This approach depends on the availability of such resources, which are very scarce. However, more sophisticated techniques have been developed that use lexical semantic knowledge to detect and handle these phenomena. For example, one interlingual approach, proposed by (Dorr, 1993), uses Jackendoff's Lexical Semantic Structure (LCS) (Jackendoff, 1983) as an interlingua. An alternative approach using lexico-structural transfer enriched with lexical semantic features that capture generalizations across the language pair was proposed by (Nasr et al., 1997). A major limitation of the interlingual and transfer approaches is that they require a large amount of explicit lexical semantic knowledge for both SL and TL.

We adopt an alternative approach called Generation-Heavy Machine Translation (GHMT) – described next. This approach is closely related to the Hybrid Natural Language Generation approach (Knight and Hatzivassiloglou, 1995; Langkilde and Knight, 1998; Bangalore and Rambow, 2000). 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 used mainly for lexical choice (including morphology and tense selection) from semantic representations (Langkilde and Knight, 1998) or from shallow unlabeled dependencies (Bangalore and Rambow, 2000).

3 Generation-Heavy Machine Translation

GHMT extends the hybrid approach to handle translation divergences *without* the use of a deeper semantic representation or transfer rules. This is accomplished through the inclusion of structural and categorial expansion of SL syntactic dependencies in the symbolic overgeneration component. The overgeneration is constrained by linguistically motivated rules that utilize TL lexical semantics and subcategorization frames and is independent of SL preferences.

Figure 1 presents an overview of the complete MT system. The three phases of Analy-

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

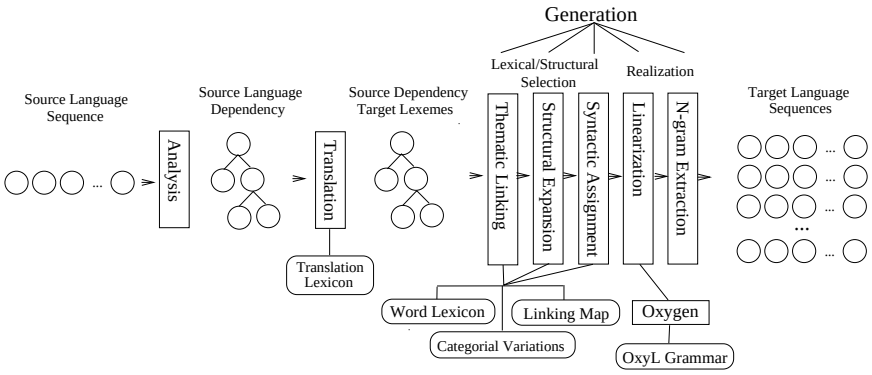

Figure 1: Generation-Heavy Machine Translation

sis, Translation and Generation are very similar to other paradigms of MT: Analysis-Transfer-Generation or Analysis-Interlingua-Generation (Dorr et al., 1999). However, Analysis and Generation in GHMT are not symmetrical. Analysis relies only on SL sentence parsing and is independent of the TL. The output of Analysis is a deep syntactic dependency in the format of the PENMAN Sentence Planning Language (SPL) (Kasper, 1989). These dependency trees normalize over syntactic phenomena such as passivization and morphological expressions (tense, number, etc.)[2]. Translation converts SL words into sets of TL words while maintaining SL dependency structure. The last phase, Generation, is where most of the work is done to manipulate the input lexically and structurally producing TL sequences. I now describe the generation component in more detail.

3.1 The Generation Component

The generation component consists of five steps (Figure 1). The first three are responsible for lexical and structural selection and the last two are responsible for realization. Initially, the SL syntactic dependency (now with TL words) is converted into a thematic dependency. This is followed by structural expansion which explores structural variations of the thematic dependency. The third step maps the thematic dependency to a target syntactic dependency. After the linearization generates a lattice of word sequences, these are ranked using a statistical n-gram model. The next section will describe the generation resources followed by a detailed explanation of the generation sub-modules.

[2]For Spanish analysis, I use the Conexor dependency parser (Tapanainen and Jarvinen, 1997) with a post-parsing step to produce the tree in the SPL format.

3.2 Generation Resources

The generation component utilizes three major resources: a word-class lexicon, a categorial-variations lexicon, and a syntactic-thematic linking map.

Word-Class Lexicon The word-class lexicon links verbs and prepositions to their subcategorization frames, thematic roles and LCS main primitives. The lexicon is organized around Levin-style classes that distinguish among different word senses. In the case of verbs, there are 511 verb classes for 3,131 verbs, totaling 8,650 entries. An example is shown here:

(1)
```
(DEFINE-WCLASS
 :NUMBER "V.13.1.a.ii"
 :NAME "Give - No Exchange"
 :SENTENCES ("He !!+ed the car to John"
             "He !!+ed John the car")
 :POS V
 :THETA_ROLES (((ag obl) (th obl) (goal obl to))
               ((ag obl) (goal obl) (th obl)))
 :LCS_PRIMS (cause go possessional)
 :WORDS (feed give pass pay peddle refund
         render repay serve))
```

In the case of prepositions, there are 43 preposition classes, for 125 prepositions, totaling 444 entries. An example is shown here:

(2)
```
(DEFINE-WCLASS
 :NUMBER "P.8"
 :NAME "Preposition Class P.8"
 :POS P
 :THETA_ROLES (time)
 :LCS_PRIMS (path temporal)
 :WORDS (until to till from before at after))
```

Note that these entries are only available for English (TL). There are no equivalent entries for any SL.

Categorial-Variation Database (CatVar) This is a database of uninflected words (lexemes) and their categorial variants[3]. The database was developed using a combination of resources and algorithms including the LCS Verb and Preposition Databases (Dorr, 2001), the Brown Corpus section of the Penn Treebank (Marcus et al., 1994), an English morphological analysis lexicon developed for PC-Kimmo (ENGLEX) (Antworth, 1990), Nomlex[4] (Macleod et al., 1998) and the Porter stemmer (Porter, 1980). The database contains 28,305 clusters for 46,037 words. The following is an excerpt:

(3)
```
(:V (hunger) :N (hunger) :AJ (hungry))
(:V (validate) :N (validation validity) :AJ (valid))
(:V (cross) :N (crossing cross) :P (across))
```

The Syntactic-Thematic Linking Map This is a large matrix extracted from the LCS Verb and Preposition Database (Dorr, 2001). It relates syntactic "cases" to thematic roles. Syntactic cases include 125 prepositions in addition to *:subj, :obj,* and *:obj2*. These are mapped to varying subsets of the 20 different thematic roles used in our system. The total number of links is 341 pairs. The following is an excerpt:

(4)
```
(:subj ->  ag instr th exp loc src goal perc poss)
(:obj2 -> goal src th perc ben)
(across -> goal loc)
(in_spite_of -> purp)
(in -> loc perc goal poss prop)
```

3.3 Thematic Linking

The first step in generation is to turn the syntactic dependency input into a thematic dependency in which all relations are thematic roles. This step deceivingly resembles SL Analysis in other MT approaches. However, it is not since the linking is applied to TL words using TL resources with no knowledge of SL preferences (except through the choice of TL words used in the translation step). This is a *loose* linking algorithm since, for example, the TL verb thematic grids are only used to determine the number and nature (obligatory, optional) of thematic roles associated with the TL verb but not how they are linked to TL syntactic positions. TL linking information is used in a later step – Syntactic Assignment. Prepositions are treated as syntactic case markers that constrain the option of thematic roles that can be assigned to their objects. For example, if a certain SL preposition is translated as the English ('to', 'toward', or 'at'), it is safe to assume that the object of the preposition is *goal* or *location* but not *source* or *purpose*.

The linking algorithm is implemented as a maximum flow network variant that uses linking constraints from the verbs and prepositions in addition to applying a Thematic Hierarchy constraint[5] and allowing all thematic roles to be treated as modifiers as a back-off option. Different linking networks are ranked with a preference for linking obligatory thematic roles over optional roles and syntactic arguments over modifiers.

Figure 2 illustrates how the correct mapping from syntax to thematic roles is done for the two sentences *Mary filled the glass with water* and *Mary filled water in the glass*. Although the second sentence is not correct English (albeit good Korean), the correct roles are assigned mainly because of the limitations imposed by allowable thematic assignments for the preposition *in* which can only link to the *goal* thematic role expected by the verb *fill*. The dotted lines in Figure 2 represent all possible links. The solid lines are the optimal linking set.

The goals of this step are (1) to reduce the number of ambiguous verb/verb-class/thematic-grid possibilities; (2) to normalize over structural variations resulting from structural and thematic divergences; and (3) to provide accurate thematic assignment that is essential for structural expansion (the next step).

3.4 Structural Expansion

This step overgenerates alternative structural configurations of the thematic dependencies. There are two operations that are applied here: Conflation and Head Swapping. Lexical-semantic information from the word-class lexicon (theta grids and lexical conceptual primitives) is used to determine the conflatability and head-swappability of combinations of nodes in the trees. Due to space limitations, I only dis-

[3] An investigation of the existence of such a resource shows that none is available. The WordNet project is currently adding such links but only for Nouns and Verbs (Christiane Fellbaum, pc.).

[4] An English Verb-Noun list extracted from Nomlex was provided by Bonnie Dorr and Greg Marton.

[5] I make an assumption here that there is a Universal Thematic Hierarchy that governs the generation of arguments. Verbs (SL or TL) that violate the Thematic Hierarchy are expected to be marked as *externalizing verbs* (Habash and Dorr, 2001).

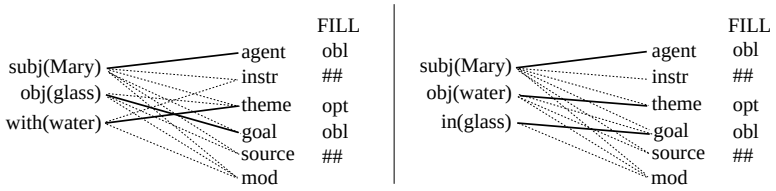

Figure 2: Syntactic-Thematic Linking Example

cuss how conflation is implemented.

For each one of the arguments of a given verb in the tree, the head verb (V_{head}) and argument (Arg) pair are checked for conflatability. A pair is conflatable if (1) there exists a verb V_{conf} that is a categorial variation of Arg (2) V_{conf} and V_{head} both share the same main lexical conceptual primitive and (3) V_{conf} can assign the same thematic roles that are assigned by V_{head} except for the role assigned to Arg. Take the following example for the Spanish *Yo le di pualadas a Juan (I gave stabs to Juan)* which results in the following thematic dependency tree after linking is done:

(5) `(3 \ |give| :ag (1 \ |I|) :th (4 \ |stab|) :goal (6 \ |Juan|))`

The theme `|stab|` has a verb categorial variation `|stab|` which belongs to two different verb classes, the Poison Verbs (as in *crucify, electrocute*, etc.) and the Swat verbs (as in *bite, claw*, etc.). Only the first class shares the same lexical conceptual primitive as the verb `|give| (CAUSE GO)`. Moreover, the verb `|stab|` requires an agent and a goal. Therefore, a conflated instance is created in this case:

(6) `(3 \ |stab| :ag (1 \ |I|) :goal (6 \ |Juan|))`

If the sentence were, say, *I gave the stab a name*, the categorial variation for `stab` would not conflate since it stands in a *goal* relationship with `give`.

3.5 Syntactic Assignment

In this step, the thematic dependency is turned into a full TL syntactic dependency. Syntactic positions are assigned to thematic roles using the verb class subcategorization frames. Different alternations associated with a single class are also generated. Class category specifications are enforced by selecting appropriate categorial variations for different arguments. For example, the main verb for the Spanish *tengo hambre (I have hunger)* translates into (have, own, possess, and be). For the last verb (be), there are different classes that have different specifications on the verb's second argument: a noun and an adjective. This, of course, results in *I am hungry* and *I am hunger* in addition to *I (have/possess/own) a hunger*. I rely on statistical extraction to decide which sequence is more likely.

3.6 Linearization

In this step a rule based linearization 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 (Habash, 2000) and makes use of the morphological generation component of the generation system Nitrogen (Langkilde and Knight, 1998).

3.7 Statistical Extraction

The final step, extracting a preferred sentence from the word lattice of possibilities is done using Nitrogen's Statistical Extractor without any changes. Sentences are scored using unigram and bigram frequencies calculated based on two years of Wall Street Journal (Langkilde and Knight, 1998).

4 Preliminary Evaluation

The following evaluation was conducted to assess the applicability of the approach to cases of Spanish-English translation divergences. The data used in the evaluation are the first 48 verb unique instances of Spanish-English divergences from the El Norte Corpus. Out of the 48 divergences, 39 (81%) were confirmed to be resolved given my approach, i.e., these divergences could be generated using the simple lexical semantics employed in GHMT together with the structural expansion and categorial variations.

On the other hand, 7 cases (14.5%) would require more conceptual knowledge. For example, the expression *dar muerte a (to give death to)*

which translates into *kill* cannot be generated currently given that in our lexicon, *kill* and *death* are not linked at all. The only verbal categorial variation of *death* is *deaden* and that is not an appropriate translation here. Generating a link between *deaden* and *kill* requires another more conceptual resource such as the Sensus Ontology (Knight and Luk, 1994). Even a simpler lexical database such as WordNet (Fellbaum, 1998) does not have a synset relating these two verbs. Such expansion is still very much in the spirit of generation-heavy machine translation since all of the new knowledge is represented in the TL.

The remaining 2 cases (4%) out of the 48 sentences require pragmatic knowledge and/or hard-wiring of idiomatic non-decompositional structures. For example the Spanish *ponerse de pie (put-self of/on foot)* should translate into *to stand up*.

5 Conclusion and Future Work

I have presented a novel MT approach that handles translation divergences between language pairs with asymmetrical resources without the use of interlinguas or structural transfer rules. Future work involves a more extensive evaluation of the Spanish-English GHMT system. The evaluation will include a test of the SL independence claim by retargeting the system to Chinese input. Extensions to both symbolic and statistical components are also planned. They include the use of conceptual representations and structural n-grams.

6 Acknowledgments

This work has been supported, in part, by ONR MURI Contract FCPO.810548265, Mitre Contract 010418-7712, DOD Contract MDA904-96-C-1250. I would like to thank Bonnie Dorr and Amy Weinberg for helpful discussions. I also would like to thank Lisa Pearl and Clara Cabezas for their help collecting and translating the Spanish evaluation data.

References

E.L. Antworth. 1990. *PC-KIMMO: A Two-Level Processor for Morphological Analysis.* Dallas Summer Institute of Linguistics.

S. Bangalore and O. Rambow. 2000. Corpus-Based Lexical Choice in Natural Language Generation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, Hong Kong, China.

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

Bonnie J. Dorr, Lisa Pearl, Rebecca Hwa, and Nizar Habash. 2002. Improved Word-Level Alignment: Injecting Knowledge about MT Divergences. In *Technical report, UMIACS-TR-2002-15, University of Maryland, College Park, MD, 2002.*

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

Bonnie J. Dorr. 2001. LCS Verb Database. Technical Report Online Software Database, University of Maryland, College Park, MD. http://www.umiacs.umd.edu/ ~bonnie/ LCS_Database_Docmentation.html.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* MIT Press. http://www.cogsci.princeton.edu/~wn.

Nizar Habash and Bonnie Dorr. 2001. Large-Scale Language Independent Generation Using Thematic Hierarchies. In *Proceedings of MT Summit VIII, Santiago de Compostella, Spain.*

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

ChungHye Han, Benoit Lavoie, Martha Palmer, Owen Rambow, Richard Kittredge, Tanya Korelsky, Nari Kim, and Myunghee Kim. 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.

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

Robert T. Kasper. 1989. A flexible interface for linking applications to PENMAN's sentence generator. In *Proceedings of the DARPA Workshop on Speech and Natural Language.* Available from USC/Information Sciences Institute, Marina del Rey, CA.

K. Knight and V. Hatzivassiloglou. 1995. 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.

K. Knight and S. Luk. 1994. Building a Large Knowledge Base for Machine Translation. In *Proceedings of AAAI-94.*

Irene Langkilde and Kevin Knight. 1998. Generation that Exploits Corpus-Based Statistical Knowledge. In *ACL/COLING 98, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*, pages 704–710, Montreal, Canada.

Benoit Lavoie, Michael White, and Tanya Korelsky. 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.

Catherine Macleod, Ralph Grishman, Adam Meyers, Leslie Barrett, and Ruth Reeves. 1998. NOMLEX: A Lexicon of Nominalizations. In *Proceedings of EURALEX'98*, Liege, Belgium.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Alexis Nasr, Owen Rambow, Martha Palmer, and Joseph Rosenzweig. 1997. 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.

M.F. Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

Pasi Tapanainen and Timo Jarvinen. 1997. A non-projective dependency parser. In *5th Conference on Applied Natural Language Processing / Association for Computational Linguistics*, Washington, D.C.