Handling Translation Divergences in Generation-Heavy Hybrid Machine Translation

Nizar Habash and Bonnie Dorr
{habash,bonnie}@umiacs.umd.edu
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20740
http://umiacs.umd.edu/labs/CLIP

Abstract

This paper describes a novel approach for handling translation divergences in a Generation-Heavy Hybrid Machine Translation (GHMT) system. The approach depends on the existence of rich target language resources such as word lexical semantics, including information about categorial variations and subcategorization frames. These resources are used to generate multiple structural variations from a target-glossed lexico-syntactic representation of the source language sentence. The multiple structural variations account for different translation divergences. The overgeneration of the approach is constrained by a target-language model using corpus-based statistics. The exploitation of target language resources (symbolic and statistical) to handle a problem usually reserved to Transfer and Interlingual MT is useful for translation from structurally divergent source languages with scarce linguistic resources. A preliminary evaluation on the application of this approach to Spanish-English MT proves this approach extremely promising. The approach however is not limited to MT as it can be extended to monolingual NLG applications such as summarization.

1 Introduction

We present a Generation-Heavy Machine Translation (GHMT) system that is asymmetrical hybrid approach to Machine Translation: our generation component constrains the translation using a combination of symbolic rules, lexicons, and corpus-based statistics. Source languages are only expected to have a syntactic parser and a translation lexicon that maps source words to target bags of words. No transfer rules or complex interlingual representations are required. The approach depends on the existence of rich target language resources such as word lexical semantics, including information about categorial variations and subcategorization frames. These resources are used to generate multiple structural variations from a target-glossed lexico-syntactic representation of the source language sentence. The multiple structural variations account for different translation divergences. The overgeneration of the approach is constrained by a target-language model using corpus-based statistics. The exploitation of target-language resources (symbolic and statistical) to handle a problem usually reserved to Transfer and Interlingual MT is useful for translation from structurally divergent source languages with scarce linguistic resources. A preliminary evaluation on the application of this approach to Spanish-English MT proves this approach extremely promising. The approach however is not limited to MT as it can be extended to monolingual NLG applications such as summarization.

The work presented here focuses on the generation component of this system 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) and (4) introduce our approach and describes the different components and algorithms in the translation system. And finally, section (5) describes a preliminary evaluation we undertook to assess the applicability of this approach on a large scale to Spanish-English MT.

2 Background

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, 1993b). 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. We will describe each divergence type before turning to alternative approaches to handling these in MT.

2.1 Translation Divergences

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. The last column displays a percentage of occurrence of the specific divergence type, taken from the first 48 verb-unique instances of Spanish-English divergences from the TREC El Norte corpus. Note that these numbers do not reflect the percentage of occurrence of the divergence type in the corpus as a whole, but rather the percentage of occurrence of the specific divergence type in the first 48 divergent sentences—and there is often overlap among the divergence types (e.g., categorial divergence occurs almost every time there is any other type of divergence).

2.1.1 Categorial Divergence

Categorial divergence involves a translation that uses different parts of speech. This is by far the most common divergence type. In the Spanish-English example below, the light verb and noun phrase are translated as another light verb and an adjectival form of the noun.

(1) tener celos (to have jealousy) ⇔ to be jealous
tener plena conciencia (have full awareness) ⇔ to be fully aware

2.1.2 Conflation

Conflation involves the translation of two words using a single word that combines the meaning of the two. In Spanish-English translation, this divergence type has two forms: light verb conflation and manner conflation. Light verb conflation involves a single verb in one language being translated using a combination of a semantically "light" verb, i.e., it carries little or no specific meaning in its own right, and some other meaning unit (perhaps a noun) to convey the appropriate meaning. English light verbs include *give*, *make*, *do*, *take*, and *have*.

(2) dar una patada (give a kick) ⇔ to kick
poner fin (put end) ⇔ to end
tomar nota (take note) ⇔ to note

Manner conflation involves translating of a single manner verb (e.g., *float*) as a light verb of motion and a manner-indicating content word that is typically a progressive manner verb in Spanish.

(3) to float ⇔ ir flotando (go (via) floating)
to pass ⇔ ir pasando (go passing)

2.1.3 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.

(4) entrar en la casa (enter in the house) ⇔ to enter the house
pedir un referendum (ask-for a referendum) ⇔ ask for a referendum

2.1.4 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.

(5) entrar corriendo (enter running) ⇔ to run in
andar volando (go-about flying) ⇔ to fly about

2.1.5 Thematic Divergence

A thematic divergence occurs when the verb's arguments switch thematic roles from one language to another. The Spanish verbs *gustar* and *doler* are examples of this case.

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

Divergence	Spanish		English	Occurrence
Categorial	X *tener hambre*	(X *have hunger*)	X *be hungry*	98%
Conflational	X *dar puñaladas a* Z	(X *give stabs to* Z)	X *stab* Z	83%
Structural	X *entrar en* Y	(X *enter in* Y)	X *enter* Y	35%
Head Swapping	X *cruzar* Y *nadando*	(X *cross* Y *swimming*)	X *swim across* Y	8%
Thematic	X *gustar a* Y	(X *please to* Y)	Y *like* X	6%

Table 1: Translation Divergence Types

(6) Me gustan uvas (to-me please grapes) ⇔ I like grapes
me duele la cabeza (to-me hurt the head) ⇔ I have a headache

2.2 Handling Translation Divergences

Since translation divergences require a combination of lexical and structural manipulation , they are traditionally minimally handled at the transfer level of the MT Hierarchy. A pure transfer approach is a brute force attempt to encode all translation divergences in a transfer lexicon (Dorr et al., 1999). However, more sophisticated techniques have been developed that use Lexical Semantic knowledge to detect and handle these phenomena. An Interlingual approach, proposed by (Dorr, 1993b; Dorr, 1994), uses Jackendoff's Lexical Semantic Structure (LCS) (Jackendoff, 1972; Jackendoff, 1976; Jackendoff, 1983; Jackendoff, 1990) as an interlingua. LCS is a compositional abstraction with language-independent properties that transcend structural idiosyncrasies. This representation has been used as the interlingua of several projects such as UNITRAN (Dorr, 1993a) and MILT (Dorr, 1997). LCS provides a granularity of representation much finer than syntactic representation and much more independent. As an example, the Spanish sentence *atravesó el río flotando* can be "composed" into the following LCS using a Spanish LCS lexicon as part of an interlingual analysis step.

(7)
```
[event CAUSE JOHN
   [event GO JOHN
      [path ACROSS JOHN
         [position AT JOHN RIVER]]]
   [manner SWIM+INGLY]]
```

In the generation phase, that same LCS is "decomposed" using LCS English lexicon entries to yield *john swam across the river*. A detailed discussion of generation from LCS is available in (Traum and Habash, 2000).

An alternative approach using lexico-structural transfer enriched with lexical semantic features was proposed by (Nasr et al., 1997). In this lexicalized grammar approach a unified syntactic and semantic representation is used for each lexical item which include appropriate cross-linguistic semantic features. Transfer lexicon rules are written as such to capture generalizations across the language pair. The transfer is done at the Deep Syntactic Structure (DSyntS) Level from Mel'cuk's Meaning Text Theory (Mel'čuk, 1988). The approach also uses Lexical Functions (also from Mel'cuk's Meaning Text Theory (Mel'čuk, 1988)) to handle analysis and generation. The following transfer rule can be used to handle the head swapping divergence discussed in the last example:

(8)
```
@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, one for English and one for Spanish. Note how the manner variable M and the path variable P switch dominance.

A major limitation of the interlingual and transfer approaches is that they require a large amount of explicit lexical semantic knowledge for both source and target languages.

3 Our Approach: Generation-Heavy Machine Translation

Our approach is closely related to the hybrid approach whose intuition was first described by the seminal work of (Knight and Hatzivassiloglou, 1995; Langkilde and Knight, 1998b; Langkilde and Knight, 1998c; Langkilde and Knight, 1998a). 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 mainly used for lexical choice (including morphology and tense selection)(Langkilde and Knight, 1998c; Bangalore and Rambow, 2000a) and for linearization from semantic representation(Langkilde and Knight, 1998a) or from shallow unlabeled dependencies(Bangalore and Rambow, 2000b).

What we propose here is the extension of the hybrid approach to handle translation divergences without the use of a deeper semantic representation or transfer rules. We accomplish this by extending the symbolic overgeneration component to include structural and categorial expansion of the source language lexico-structural representation. By overgenerating lexico-structural combinations preferred by the target language, we make them available choices for ranking by the statistical extraction component. the overgeneration is constrained by linguistically motivated rules that utilizes target language lexical semantics and subcategorization frames and is independent of the source language preferences.

3.1 Overview of GHMT

Figure (1) presents an overview of the complete MT system. The three phases of Analysis, Translation and Generation are very similar to other paradigms of MT (Analysis-Transfer-Generation or Analysis-Interlingua-Generation)(Dorr et al., 1999). However, these phases are not symmetrical. Analysis relies only on the source-sentence parsing and is independent of the target language. 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 source-language lexemes into bags of target-language lexemes. The dependency structure of the source language is maintained. The last phase, Generation, is where most of the work is done to manipulate the input lexically and structurally produce target sequences.

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

3.1.1 Word-Class Lexicon

The word-class lexicon currently contains only verbs and prepositions, as these are the predicate-argument relations primarily involved in translations—each of these categories are grouped into "classes." In the case of verbs, there are 511 verb classes for 3,131 verbs, totaling 8,650 entries. An example is shown here:

```
(9) (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)
     :SPEC ((ag (animate +)))
     :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:

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

Note that these entries are only available in the system for English since it is the target language. There are no equivalent entries for the source language.

3.2 Categorial-Variation Database

The Categorial-Variation Database (CatVar) is a database of words and their categorial variants. Our investigation of the existence of such a resource so far shows that none is available.[2] Our research has involved the creation of resource for English. The structure of this

[2] The WordNet project is currently adding such links but only for Nouns and Verbs (Christiane Fellbaum, pc.).

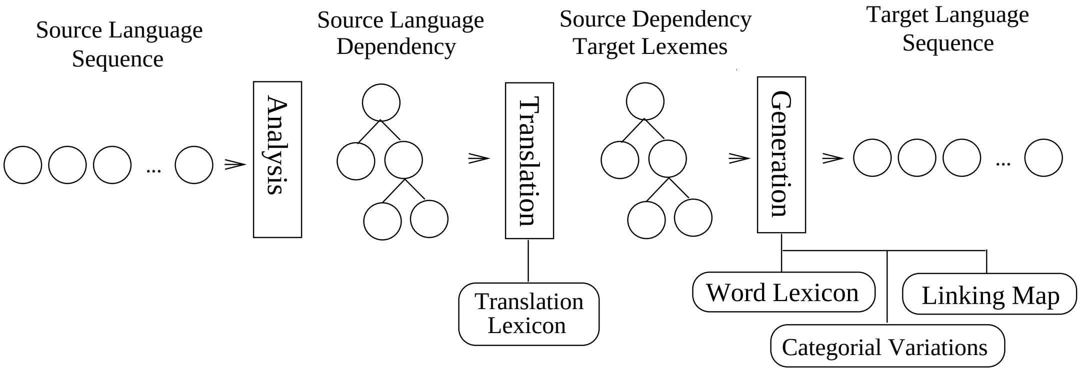

Figure 1: Generation-Heavy Machine Translation

database is rather simple: it is flat with an indexing file that is accessible through a hash table. For a given word and its optional parts of speech, the lookup mechanism returns a list of lists of categorial variants of the word (including the word itself). An excerpt is shown here:

(11)
```
(:V (hunger) :N (hunger) :AJ (hungry))
(:V (validate) :N (validation validity)
 :AJ (valid))
(:V (cross) :N (crossing cross)
 :P (across))
```

We have currently developed 28,305 catvar clusters for 40,443 POS sub-cluster, totaling 46,037 words (lexemes). The database was developed using a combination of resources and algorithms including the LCS Verb and Preposition Databases (Dorr, 2001b; Dorr, 2001a), the Brown Corpus section of the Penn Treebank (Marcus et al., 1994), an English morphological analysis lexicon developed for PC-Kimmo (ENGLEX) (Antworth, 1990), and the porter stemmer(Porter, 1980).

3.3 The Syntactic-Thematic Linking Map

This is a large matrix that was extracted from the LCS Verb Database (LVD)(Dorr, 2001b) and the LCS Preposition Database (LPD)(Dorr, 2001a). It relates syntactic "cases" to thematic roles. Thematic 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. An excerpt of this resource is shown below.

(12)
```
(:subj ag instr th exp loc src goal perc
       mod-poss poss)
(:obj2 goal src th perc ben)
(aboard loc goal)
(about info mod-perc perc poss time purp
       loc goal pred)
(according_to purp)
(across goal loc)
(in_spite_of purp)
(in loc mod-poss perc goal poss prop)
```

4 The Generation Component

The input to the generation component is a deep syntactic dependency tree similar to Mel'cuk's Meaning Text Theory (MTT) (Mel'čuk, 1988), but it is written in the format of the PENMAN Sentence Planning Language (SPL) (Pen, 1989). The part-of-speech and roles definitions are very small. There are 10 parts of speech (verb, preposition, noun, proper noun, adjective, adverb, determiner, conjunction, interjunction and punctuation) and only 4 roles, Subject, object, indirect object (which map to I,II, and III in MTT) and modifier. All nodes in the dependency tree are expected to be ambiguous bags of lexemes. Our machine translation approach involves a lexical translation of the parse-tree nodes corresponding to words in the source-language sentence. No structural transfer is required.

4.1 Thematic Linking

The first step in our system is to turn the syntactic dependency input into a thematic dependency tree. The syntax-thematic linking here is achieved through the use of thematic grids associated with English (verbal) head nodes. This step is done in the generation process using the target-language resources only. Therefore, it is a *loose* linking algorithm that is constrained by

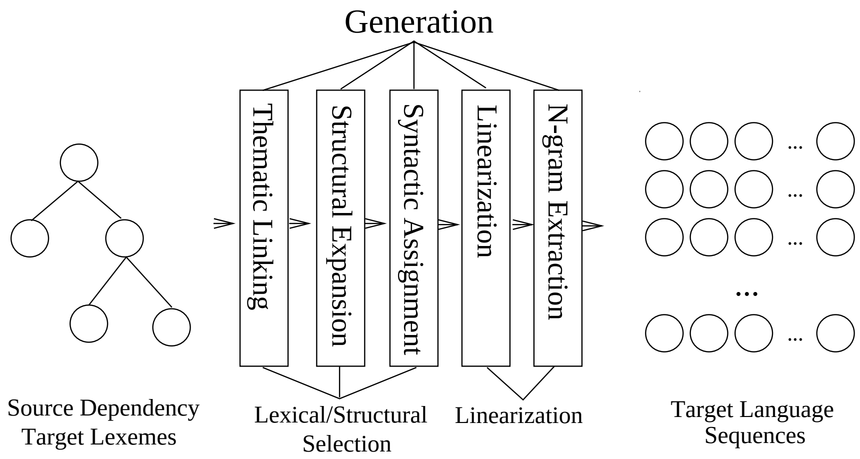

Figure 2: Generation

the thematic grids of the predicates in the target language (verbs and prepositions).

Prepositions are treated as syntactic case markers that constrain the option of thematic roles that can be assigned to their objects. The number and nature (obligatory, optional) of the thematic roles are determined by the verb thematic grid. We treat the linking problem as a maximum flow network variant that uses linking constraints from the verbs and prepositions in addition to applying a Thematic Hierarchy constraint[3] and allowing all syntactic roles to be treated as modifiers as an option. Therefore, we are guaranteed to get a network in every case. However, the different resulting networks are ranked by a criterion that prefers obligatory thematic roles to be linked, prioritizing linkings involving arguments ahead of those involving modifiers.

Figure (3) 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 example is not correct English (albeit good Korean), the correct roles are assigned mainly because of the limitations imposed by allowable thematic assignments for the prepositions.

The output of this phase is a thematic dependency in which the relations of children to parents are thematic roles (and modifiers) instead of syntactic roles. The goals of this phase are many: 1) Reduction of ambiguity. Since each verb can have multiple verb class memberships (some of which have multiple thematic grids), this step reduces the verb/verb-class/grid possibilities to only those that rank highest according to the criteria described earlier. 2) Normalization: This step normalizes over structural variation and thus approaches a solution to the structural and thematic divergences on a thematic level.[4] 3) Accurate thematic assignment, which is essential for handling structural variations.

This step looks like analysis but it is fully driven by resources and constraints from the target language. That is why it is a central step in this generation-heavy approach.

4.2 Structural Expansion

This step is for exploring alternative structural configurations of the input. There are two operations that are applied here: Conflation and Head Swapping. Lexical-semantic information from the verb class lexicon (both theta grids and lexical conceptual primitives) is used to determine the conflatability and head-swappability of combinations of nodes in the trees.

4.2.1 Conflation

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

[3] We make an assumption here that there is a Universal Thematic Hierarchy that governs the generation of arguments. Predicates that violate the thematic hierarchy are expected to be marked as externalizing predicates in both source and target languages(Habash and Dorr, 2001)

[4] This also applies to expanding the possible set of alternations eventually.

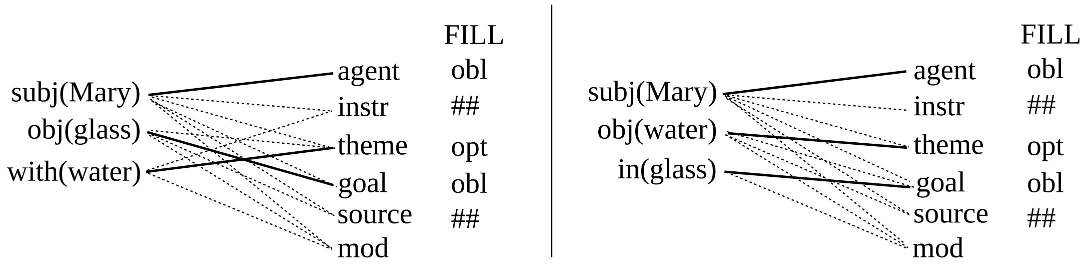

Figure 3: Syntactic-Thematic Linking Example

is conflatable if (1) there exists a verb V_{conf} that is a categorial variation of the argument (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 one that is 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:

```
(13) (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 for the stabbing. Therefore, a conflated instance is created in this case:

```
(14) (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 have failed because it stood in a goal relationship with its parent.

4.2.2 Head Swapping

Unlike Conflation, Head Swapping is restricted to head-modifier pairs. Every such pair's swappability is determined by the following criteria: (1) there exists a verb V_{conf} that is a categorial variation of the modifier (2) there is a categorial variation of V_{head} that can become a child of V_{conf} such as a noun, adjective, adverb or even a preposition. (3) all the arguments before the swapping retain their thematic roles regardless of whether they move with the swapped verb or not. For example, the German *ich esse gern (I eat likingly)* results in the following thematic dependency tree after linking is done:

```
(15) (3 \ |eat|
        :th (1 \ |I|)
        :mod (6 \ |like|))
```

Here the modifier `|like|` and the main verb `|eat|` can be swapped to produce *I like eating* or *I like to eat*. If the demoted verb can become a preposition, the swapping is more complicated since prepositions are not part of the thematic dependency. For example, the Spanish *Juan cruzó el río nadando (Juan crossed the river swimming)* results in the following thematic dependency tree after linking is done:

```
(16) (3 \ |cross|
        :th (1 \ |Juan|)
        :loc (4 \ |river|)
        :mod (6 \ |swim|))
```

The modifier `|swim|` is itself a verb. And the main verb `|cross|` has a prepositional categorial variation `|across|` which can assign the thematic role `:loc` to `|river|`:

```
(17) (3 \ |swim|
        :th (1 \ |Juan|)
        :mod (4 \ |river| :prep |across|))
```

4.3 Syntactic Assignment

In this step, the thematic dependency is turned into a full target syntactic dependency. Syntactic positions are assigned to thematic roles using the verb class subcategorization frames. Different alternations associated with a single class are generated here too which allows for a widening range of expression that is specific to the target language. Class category specifications are enforced by picked appropriate categorial variations of the 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*. That is where statistical extraction is most valuable; to decide which sequence is more likely.

4.4 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, 1998b). The grammar is based on previous work we have done in Chinese-English LCS-based MT(Dorr et al., 1998; Traum and Habash, 2000).

4.5 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, 1998c).

5 Preliminary Evaluation

We conducted the following evaluation to assess the applicability of the approach on handling Spanish-English translation divergences. The data we use for our evaluation is the first 48 verb unique instances of Spanish-English variations from the El Norte Corpus. Out of the 48 sentences, 39 (81%) were confirmed to be resolved given our approach, i.e., these divergences could be generated using the simple lexical semantics we employ 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 target language.

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*.

6 Future Work

Our immediate future work will involve an expansion of the linearization grammar to be able to handle large-scale Spanish-English GHMT. We also plan to explore extensions to the symbolic component of our system, e.g., a conceptual representation that facilitates generation by linking concepts that are not related morphologically. In addition, we plan to explore extensions to the statistical component through the use of structural bigrams. And finally, we are interested in testing our source-language independence claim by retargeting the system to Chinese input.

7 Acknowledgments

This work has been supported, in part, by ONR MURI Contract FCPO.810548265, DARPA TIDES Contract N66001-00-2-8910, and DOD Contract MDA904-96-C-1250. We would like to thank Lisa Pearl and Clara Cabezas for their help collecting and translating the Spanish data for our evaluation. We would also like to thank Amy Weinberg for helpful conversations.

References

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

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

S. Bangalore and O. Rambow. 2000b. Exploiting a probabilistic hierarchical model for generation.

Bonnie J. Dorr, Nizar Habash, and David Traum. 1998. A Thematic Hierarchy for Efficient Generation from Lexical-Conceptal Structure. In *Proceedings of the Third Conference of the Association for Machine Translation in the Americas, AMTA-98, in Lecture Notes in Artificial Intelligence, 1529*, pages 333–343, Langhorne, PA, October 28–31.

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 *submitted to the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA.

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

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

Bonnie J. Dorr. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20(4):597–633.

Bonnie J. Dorr. 1997. Large-Scale Acquisition of LCS-Based Lexicons for Foreign Language Tutoring. In *Proceedings of the ACL Fifth Conference on Applied Natural Language Processing (ANLP)*, pages 139–146, Washington, DC.

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

Bonnie J. Dorr. 2001b. 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 [2000, September 7].

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.

Ray Jackendoff. 1972. Grammatical Relations and Functional Structure. In *Semantic Interpretation in Generative Grammar*. The MIT Press, Cambridge, MA.

Ray Jackendoff. 1976. Toward an Explanatory Semantic Representation. *Linguistic Inquiry*, 7(1):89–150.

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

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

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. 1998a. Generating Word Lattices from Abstract Meaning Representation. Technical report, Information Science Institute, University of Southern California.

Irene Langkilde and Kevin Knight. 1998b. 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.

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

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.

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

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.

ISI, University of Southern California, 1989. *The Penman Reference Manual*, December.

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

David Traum and Nizar Habash. 2000. 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.