Interlingua Approximation: A Generation-Heavy Approach

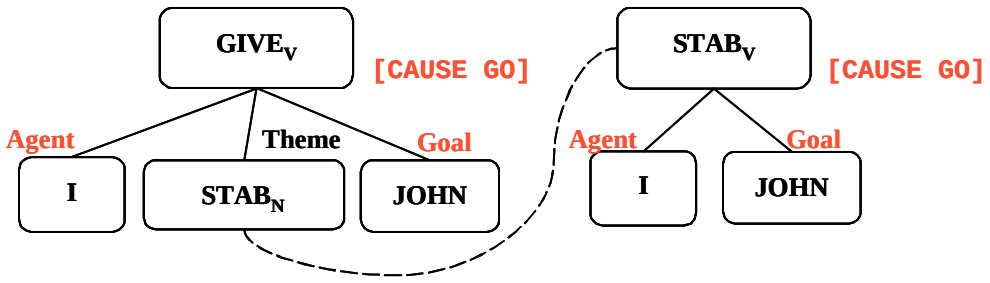
Bonnie Dorr and Nizar Habash
University of Maryland
Institute for Advanced Computer Studies
{bonnie,habash}@umiacs.umd.edu

1. Introduction

To date, construction of interlingual resources continues to be a labor-intensive process—often resulting in knowledge-based systems that suffer from a lack of robustness. Such systems may work well on certain types of phenomena, but their complex knowledge-based foundation makes them difficult to extend to new phenomena or languages. We adopt the view that it is possible to *approximate* the depth of knowledge-based interlingual systems by tapping into the richness of target-language (TL) resources (i.e., English, in our projects) and using this information to map the source-language (SL) input to the English output. A key feature of our approach is the use of some, but not all, components of an interlingual representation (e.g., the top-level primitives and basic argument structure) to map representations associated with a resource-poor language into those of a resource-rich language. The approach lends itself to the generation of multiple sentences that are statistically pared down so that the most likely sentence is generated according to the constraints of the TL.

Consider the oft-cited Spanish example, "Yo le di puñaladas a John" (I gave knife-wounds to John, i.e., "I stabbed John"). Such cases are traditionally handled in interlingual systems by means of decomposition into a conceptual representation (Dorr, 1993). We espouse a more economical approach that uses the structure of syntactic dependencies coupled with knowledge encoded in the Lexical Conceptual Structure Verb Database (LVD) of (Dorr, 2001).

More specifically, rather than mapping the SL input into a representation with the full range of interlingual components, this simpler approach uses only the argument structure of the input dependency tree and top-level conceptual nodes (such as the "CAUSE GO") coupled with thematic-role information. In order to produce a TL (English) sentence from this representation, the top-level conceptual nodes are first checked for possible matches—and then conflated arguments (the **STAB$_N$** node below) are potentially absorbed into other predicate positions, as long as there is a relation between the conflated argument and the new predicate node, disregarding part-of-speech (in this case **STAB$_V$**). This process is shown pictorially below.



Note that there is nothing inherent in the design of this matching approach that would prevent the sentence "I gave a stab to John" from being generated. However, we rely on a downstream statistical extraction module to rank all possible outputs of the module, e.g., "I gave a stab to John" is ranked lower than the more preferred sentence "I stabbed John". This approach—which we call Generation-Heavy Hybrid Machine Translation (GHMT; Habash and Dorr, 2002)—produces a list of sentences with their associated ranking. In the example given above, the output of our system is as follows:

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

The remainder of this paper describes the resources and processes associated with the GHMT approach. We will demonstrate the usefulness of our approach as an approximation to interlingual approaches—showing that it is possible to accommodate cross-lingual phenomena without significant engineering of SL resources.

2. GHMT

GHMT exploits deep TL resources to translate from SLs with shallow resources. The approach relies on the pre-existence of a SL dependency parser and a bilingual translation lexicon (i.e., a "tralex"). TL resources include a simplified version of LVD representations (both top-level primitives and thematic roles) and a database of categorial variation classes (the "CatVar" database built in-house at University of Maryland).[1] TL resources are used to overgenerate structural variations of SL dependency trees instead of depending on transfer rules or interlingual representations. These same resources are used to constrain the ambiguity resulting from lexical transfer. There are multiple advantages to this "approximate interlingua" approach over "full interlingua" or transfer approaches, including shorter development time and SL independence.

Most interlingual (IL) representations include: primitives, relations, bi-directional lexicons mapping between IL and a Human Language (HL). During analysis, the IL representation is constructed (composed) from the SL input using the SL lexicon. Then the generation step creates the TL output is created by deconstructing (decomposing) the IL using the TL lexicon. In principle, this process is completely symbolic and only dependent on the SL and TL lexicons as in LCS-based MT (Dorr, 1993). Nitrogen (Langkilde and Knight, 1998) provides a hybrid approach to generation from AMR by overgenerating and integrating a statistical language model to rank possible sequences.

[1] The categorial variation database contains 28K classes of related lexemes, covering a total of 46K English lexemes; cf. (Habash, 2002).

The GHMT approach borrows this same statistical ranking scheme from Nitrogen, but it is designed to approximate the definition and behavior of an IL approach to handling MT divergences. The primary distinction between GHMT and other approaches is its incorporation of alternatives to traditional primitives, semantic relations, and lexical information. We describe each of these in turn, identifying places where our approach differs from other primitive-based models such as LCS-based MT and AMR-based generation.

2.1. GHMT Primitives

Primitives in the IL are the primary units of meaning specification. Their granularity and their relation to lexemes in HL differ from one IL to another. Although the concept of a primitive is theoretically simple, its implementation is very hard. Researchers developing IL representations rarely agree on the form or even the meaning of a primitive. Lexical Conceptual Structure (LCS) distinguishes between closed-class primitives that are general meaning specifiers such as **GO**, **BE** and **CAUSE** and open class primitives such as **OPEN+INGLY** and **HUNGER+INGLY**. Abstract Meaning Representation (AMR) uses a large set of hierarchically related primitives from the Sensus Ontology (Knight and Luk, 1994) such as

```
|status,condition|
     |physiological state|
          |hungriness|
               |malnourishment|
```

In addition, both LCS and AMR include special closed-set features to represent sub-lexical information such as tense, number, gender and part of speech.

In such primitive-based approaches, it is difficult to decide what a primitive token should denote. In general, it is assumed that $hunger_N$, $hunger_V$, and $hungry_{Adj}$ share a primitive specifying their common concept, but deciding what that basic concept should be is non-trivial (e.g., the state **HUNGRY** vs. the condition **HUNGER**). The selected concept must be present in the definition of other words related to *hunger* such as English *starvation*, Spanish *hambre*, and Arabic جوع . The question is whether it is possible to define language-independent primitives with enough granularity to disallow the expansion/alteration of the meaning of a lexeme when mapping it into an IL concept.

This question is addressed in GHMT by the use of three resources: the SL-TL bilingual translation lexicon (tralex), the categorial variation database (CatVar), and the statistical language model. The tralex expands the SL word into a set of TL words (e.g. hambre → {famine hunger starvation}). The CatVar allows expansion of any of these TL words to their other parts of speech (e.g. $starvation_N$ → $starve_V$). The TL statistical model ranks the different expansions in their contexts to select the most likely TL sequences.

2.2. GHMT Semantic Relations

Whereas primitives specify the *content* of the IL, semantic relations are the primary units of *structure* in the IL. Semantic relations are defined differently in various ILs but they are always expected to normalize over syntactic (surface) structure variations while maintaining the logical relationship between different contents. The representation of

semantic relations is a very complex problem since there are so many linguistic phenomena such as verb alternations and transformations that need to be handled consistently both monolingually and translingually. In systems where deep representations are used, the granularity of semantic relations may be very high. In order to provide broad-coverage of linguistic phenomena, such systems rely crucially on complex (and thus expensive) lexicons.

In contrast, GHMT handles relations at the thematic level using a combination of algorithms (thematic linking and structural expansion), TL resources (statistical language models) and the language-independent principle of Universal Thematic Hierarchy (UTH). Unlike IL systems that expect the analysis step to select the thematic representations for the SL predicates, thematic linking is fully handled at the TL side as part of the generation step. GHMT relies on TL subcategorization frames, which specify the thematic roles licensed by verbs and prepositions to assign thematic roles to the translated SL words. No strict matching is enforced here, so non-English structures such as *John filled the water in the glass* may still be thematically linked as **`(fill :agent John :theme water :location glass)`** since the preposition *in* assigns *location* to its object. A later (stricter) syntactic-assignment step only allows TL configurations to be generated. Finally, thematic relations are further refined during a structural-expansion stage which explores conflation and head-swapping manipulations of the thematically linked SL structures.

Once the thematic relations are established, two statistical models (surface n-grams and structural lexemal n-grams) are used to select among the overgenerated structures. As for thematically divergent verbs such as Psych verbs, e.g. *like vs. please*, these are required to be marked in the lexicons as externalizing verbs that violate the Universal Thematic Hierarchy not as violators of how their language behaves relative to another language.

2.3. GHMT Lexicons

In traditional primitive-based approaches, the SL and TL lexicons specify lexical entries that map between surface words and relations into IL primitives and relations. For an IL to behave properly, the SL and TL lexicons should be symmetric in their coverage—an expensive task to accomplish. In GHMT, symmetry of resources is not required, as long the TL side is very rich. The current implementation of GHMT incorporates rich TL (English resources), including verb subcategorization frames, categorial variations, lexical-conceptual information (LCS-based) in addition to the two statistical language models mentioned above. In principle, any already existing resource of HL that was designed with the depth of an IL can be used within this approach to build a GHMT system targeting this HL.

3. MT Divergence Handling

Translation divergences occur when the gist/meaning of a sentence is spread over different words and relations from one language to another. Translation divergences need to be handled at the transfer and IL level of the MT hierarchy because they require a lot of structural manipulations. What makes divergences quite hard for transfer systems

is that the different divergence types can co-occur which means every combination needs to be listed in a transfer lexicon. On the other hand, an IL with the proper granularity such as LCS can provide a consistent simple representation for handling translation divergences. However, the traditional LCS-based approach requires a great deal of resources on both the SL and TL sides. The table below illustrates how certain divergence types are handled in LCS-based MT and GHMT. More details are in (Habash 2002) and (Habash and Dorr 2002).

Divergence Type	LCS-based MT	GHMT
Categorial	SL and TL lexical entries with shared primitives	Categorial variation database
Conflational	SL and TL lexical entries shared substructures	Structural expansion
Structural	SL and TL lexical entries with argument-position markings	Thematic linking, syntactic assignment
Head-swapping	TL lexical entries marked for predicate reversal w.r.t SL predicates	Structural expansion
Thematic	TL lexical entries marked for argument reversal w.r.t. SL arguments	Thematic linking, universal thematic hierarchy

The bottom line is that most traditional IL approaches do not explain certain language behaviors that appear to be statistical in nature—the prototypical case of which was presented earlier in the "stab" example. Moreover, most IL approaches are analysis-focused, thus requiring a significant investment of effort in SL development, including in cases where the SL resources are rare or non-existent. The GHMT approach provides a solid foundation for re-use of already existing components for MT from new languages, a characteristic we've labeled "re-source-ability" (as opposed to "re-target-ability"). Finally, the same generation-heavy paradigm can be employed for correcting and expanding generated phrases that are monolingually induced, e.g., alternations described in (Levin, 1993), such "I stuffed socks in the drawer" vs. "I stuffed the drawer with socks."

Acknowledgements

This work has been supported by Mitre Contract 010418-771 and ONR Muri Contract FCPO.810548265.

References

Dorr, Bonnie J. 2001. LCS Verb Database, Online Software Database of Lexical Conceptual Structures and Documentation, UMCP.
http://www.umiacs.umd.edu/~bonnie/LCS_Database_Documentation.html

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

Habash, Nizar. 2002. Generation-Heavy Hybrid Machine Translation, Proceedings of the Second International Natural Language Generation Conference, New York.

Habash, Nizar and Bonnie Dorr. 2002. Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Machine Translation, Proceedings of AMTA, Tiburon, CA.

Hovy, E.H. 1998. Combining and Standardizing Large-Scale, Practical Ontologies for Machine Translation and Other Uses. Proceedings of the International Conference on Language Resources and Evaluation (LREC). Granada, Spain.

Knight, K. and S.K. Luk. 1994. Building a Large-Scale Knowledge Base for Machine Translation. Proceedings of the AAAI Conference.

Langkilde, Irene and Kevin Knight. 1998. Generation that Exploits Corpus-Based Statistical Knowledge, Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (joint with the 17th International Conference on Computational Linguistics), Montreal, Canada, 704-710.

Levin, Beth. 1993. English Verb Classes and Alternations: A Preliminary Investigation, University of Chicago Press, Chicago, IL.