A Multi-Path Architecture for Machine Translation of English Text into American Sign Language Animation

Matt Huenerfauth
Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
matthewh@seas.upenn.edu

Abstract

The translation of English text into American Sign Language (ASL) animation tests the limits of traditional MT architectural designs. A new semantic representation is proposed that uses virtual reality 3D scene modeling software to produce spatially complex ASL phenomena called "classifier predicates." The model acts as an interlingua within a new multi-pathway MT architecture design that also incorporates transfer and direct approaches into a single system.

1 Introduction and Motivation

American Sign Language (ASL) is a visual/spatial natural language used primarily by the half million Deaf individuals in the U.S. and Canada. ASL has a distinct grammar, vocabulary, and structure from English, and its visual modality allows it to use linguistic phenomena not seen in spoken languages (Liddell, 2003; Neidle et al., 2000). English-to-ASL translation is as complex as translation between pairs of written languages, and in fact, the difference in modality (from a written/spoken to a visual/spatial manually performed system) adds new complexities to the traditional MT problem.

Building an English-to-ASL MT system is important because although Deaf students in the U.S. and Canada are taught written English, the difficulties in acquiring a spoken language for students with hearing impairments prevents most Deaf U.S. high school graduates from reading above a fourth-grade level (students age 18 and older reading text at a typical 10-year-old level) (Holt, 1991). Unfortunately, many Deaf accessibility aids (e.g. television closed captioning or teletype telephone services) assume that the viewer has strong English literacy skills. Since many of these individuals are fluent in ASL despite their difficulty reading English, an ASL MT system could make more information and services accessible in situations where English captioning text is above the reading level of the viewer or a live English-to-ASL interpreter is unavailable.

Researchers in graphics and human figure modeling have built animated models of the human body that are articulate enough to perform ASL that native signers can understand (Wideman and Sims 1998). Most animation systems use a basic instruction set to control the character's movements; so, an MT system would need to analyze an English text input and produce a "script" in this instruction set specifying how the character should perform the ASL translation output. The MT task is conceived of as translation from English text into this script because ASL has no written form. While linguists use various ASL glosses, all were designed to facilitate linguistic study, not to serve as a natural writing system, and so they omit certain details.

Since there is no ASL orthography used by the Deaf community, there are no natural sources of ASL corpora. To collect a corpus for statistical MT research, a movement annotation standard must be developed, ASL performances videotaped, and finally the videos manually transcribed – a slow and expensive process (Niedle, 2000). Motion-capture glove technology may seem like a solution to this problem, but this type of data cannot easily be synthesized into novel and fluent ASL animations. The difficulty in obtaining large corpora of ASL is why statistical approaches to the English-to-ASL MT problem are not currently practical.

2 ASL Linguistic Issues

As opposed to spoken/written languages, ASL relies on the multiple simultaneous channels of handshape, hand

location, palm orientation, hand/arm movement, facial expressions, and other non-manual signals to convey meaning. To express additional meaning, ASL may modify aspects of the manual performance of a sign (handshape, timing, motion path, repetition, etc.), perform an additional grammatical facial expression, or systematically use the areas of space around the signer.

ASL signers use the space around them for several grammatical, discourse, and descriptive purposes. During a conversation, an entity under discussion (whether concrete or abstract) can be "positioned" at a point in the signing space. Subsequent pronominal reference to this entity can be made by pointing to this location, and some verb signs will move toward or away from these points to indicate their arguments. Generally, the locations chosen for this pronominal use of the signing space are not topologically meaningful; that is, one imaginary entity being positioned to the left of another in the signing space doesn't necessarily indicate the entity is left of the other in the real world.

Other ASL expressions are more complex in their use of space and position invisible objects around the signer to topologically indicate the arrangement of entities in a 3D scene being discussed. Special ASL constructions called "classifier predicates" allow signers to use their hands to represent an entity in the space in front of them and to position, move, trace, or re-orient this imaginary object in order to indicate the location, movement, shape, or other properties of some corresponding real world entity under discussion. A classifier predicate generally consists of the hand in one of a closed set of semantically meaningful shapes as it moves in a 3D path through space in front of the signer.

For example, the sentence "the car drove down the bumpy road past the cat" could be expressed in ASL using two classifier predicates. First, a signer would move a hand in a "bent V" handshape (index and middle fingers extended and bent) forward and downward to a point in space in front of his or her torso where an imaginary miniature cat could be envisioned. Next, a hand in a "3" handshape (thumb, index, middle fingers extended) could trace a path in space past the "cat" in an up-and-down fashion as if it were a car bouncing along a bumpy road. Generally, "bent V" handshapes are used for animals, and "3" handshapes, for vehicles.

The ability of classifier predicates to topologically represent a three-dimensional scene make them particularly difficult to generate using traditional computational linguistic methods and models. To produce this pair of classifier predicates, there must be a spatial model of how the scene is arranged including the locations of the cat, the road, and the car. A path for the car must be chosen with beginning/ending positions, and the hand must be articulated to indicate the contour of the path (e.g. bumpy, hilly, twisty). The proximity of the road to the cat, the plane of the ground, and the curve of the road must be selected. Other properties of the objects must be known: (1) cats generally sit on the ground and (2) cars usually travel along the ground on roads. The successful translation of the English text into these classifier predicates used a great deal of semantic analysis, spatial knowledge, and reasoning.

3 ASL MT Architectural Designs

There is an architectural spectrum along which most MT systems can be classified; loosely they are grouped into three basic designs: direct, transfer, or interlingua (Dorr et al., 1998). Direct systems process individual words of the source language text; translation is achieved without performing any syntactic analysis. Transfer systems do analyze the input text to some syntactic or semantic level, and then a set of "transfer" rules produce a corresponding syntactic or semantic structure in the target language. Finally, a generation component converts this structure into a target-language text. Interlingual systems take this analysis of the input text one step further: the source is analyzed and semantically processed to produce a typically language-independent semantic representation called an "interlingua," and then a generation component produces the target-language surface form from there. These design choices are often pictured as a pyramid, as in Figure 1, adapted from a figure in (Dorr et al., 1998).

Generally, in the absence of statistical or case-based information, the higher up the pyramid that the source text is analyzed, the more complex and subtle are the divergences the system can handle. In particular, at the interlingual level, a knowledge base can supplement the linguistic information, producing translations that use world knowledge and that may convey more information than was present in the source text (devoid of context). However, any of the approaches can produce a correct translation for certain inputs since not all sentences require such sophisticated analysis to be translated – some exhibit little translation divergence. Another trend as one goes up the MT pyramid is that the

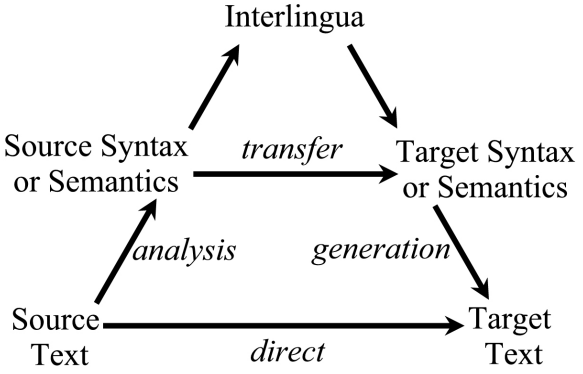

Figure 1: Pyramid of MT Architecture Designs.

amount of domain specific development work that must be performed increases dramatically. While direct systems may only require a bilingual lexicon, transfer systems also require analysis and transfer rules. Interlingual systems require interlingual representations and sometimes domain specific knowledge bases.

Non-statistical direct approaches to English-to-ASL MT generally produce simple translations that are often little more than word-to-sign dictionary look-ups. With the addition of some basic sentence reordering heuristics, such systems can occasionally produce acceptable output on simple English inputs or on those English-ASL sentence pairs that have similar word order.[1] Since no syntactic analysis is performed, there is no chance that an input sentence will be outside the linguistic coverage of the system; so, the translation process will always produce some output. Even if an English word is not in the translation lexicon, manual fingerspelling can be used to express the word.

Transfer MT designs address most of the linguistic shortcomings of direct systems but do require additional linguistic resources to be developed. There have been a few transfer-based English-to-ASL systems built (Huenerfauth, 2003), and several have had success in particular aspects of the MT task, like expressing adverbials (Zhao et al., 2000) or representing ASL phonological information (Speers, 2001; Sáfár and Marshall, 2001). These systems show promise that a transfer approach could someday handle most ASL sentences that do not require complex or topological use of the signing space. As the "bumpy road" example illustrates, generating classifier predicates would require more than a simple syntactic or semantic analysis – spatial analogy, scene visualization, and/or some degree of iconicity seem to be involved.[2]

For this reason, ASL transfer systems merely omit classifier predicates from their coverage; however, many English concepts lack a fluent ASL translation without them. Further, these predicates are common in ASL; signers generally produce a classifier predicate at least once per minute (once per 100 signs) (Morford and MacFarlane, 2003). So, systems that cannot produce classifier predicates are not a viable long-term solution to the English-to-ASL MT problem. To supply the semantic understanding, spatial reasoning, and world knowledge that classifier predicate generation demands, an interlingual approach (one with deeper semantic analysis and 3D spatial representations) is required.

[1] Direct systems more readily convert English text into a signing system like Signed Exact English, a manually coded form of English, not a distinct natural language, like ASL.

[2] Linguists debate whether classifier predicates are paralinguistic iconic gestures, non-spatial polymorphemic constructions, or compositional yet spatially-aware expressions (Liddell, 2003), but transfer approaches to MT seem ill-suited to producing classifier predicates in any case.

4 A Multi-Path MT Architecture

While an interlingual approach to the classifier predicate translation task sounds useful, there is a problem. It's hard to built a true interlingual system for anything but a carefully limited domain; building the linguistic and knowledge resources needed for interlingual translation on less restricted texts can entail too much overhead to be practical. What is special about the MT problem for ASL – and the reason why interlingual translation may be possible – is that we can characterize and identify the "hard" input sentences, the ones that require classifier predicates for translation. These are spatially descriptive English input texts, those generally containing: spatial verbs describing locations, orientations, or movements; spatial prepositions or adverbials with concrete or animate entities; or lexical items related to other common topics or genres in which classifier predicates are typically used. Such genres (e.g. vehicle motion or furniture arrangement in a room) could be detected using the features mentioned above.

While an interlingual approach is needed to translate into classifier predicates, there are a vast number of English input sentences for which such deep analysis and reasoning would not be necessary. As we've seen from the direct and transfer discussion above, these resource-lighter approaches can often produce a correct translation from lexical or syntactic information alone.

This analysis suggests a new multi-path architecture for an MT system – one that includes a direct, a transfer, and an interlingual pathway. English input sentences within the implemented interlingua's limited domain could follow that processing pathway, those sentences outside of the interlingual domain but whose syntactic features fall within the linguistic coverage of the analysis and transfer rules could use the transfer pathway, and all other sentences could use the direct pathway with its bilingual dictionary look-up.

Limiting the domain that the transfer and interlingua components must handle makes the development of these components more manageable. The transfer pathway's analysis grammar and transfer rules would not have to cover every possible English sentence that it encounters: some sentences would simply use the direct translation pathway. Limiting domains has an even more dramatic benefit for the interlingual pathway. Instead of building interlingual analysis, representation, and generation resources for every possible domain, the interlingual development can focus on the specific domains in which classifier predicates are used: walking upright figures, moving vehicles, furniture or objects arranged in a room, giving directions, etc. In this way, the "depth" of divergence-handling power of some translation approaches and the "breadth" of coverage of others can both be part of this multi-path architecture.

This design does more than just restrict the domains for which the interlingua must be implemented; it also reduces the ontological complexity that the entire interlingua must support. The domains listed above share a common feature: they all discuss the movement, location, orientation, and physical description of entities in three-dimensional scenes. Some complex phenomena whose handling often makes designing an interlingual representation quite difficult – abstract concepts, beliefs, intentions, quantification, etc. – do not need to be represented. In a sense, this multi-path architecture doesn't just limit the things that must be represented, but the "type" of these things as well.

Having multiple processing pathways does not mean that there is necessarily a new problem of choosing which to use. The system could be implemented as a 'fall back' architecture in which the system could attempt the most complex approach (interlingual) and drop back to each of the simpler approaches whenever it lacks the proper lexical, syntactic, semantic, or knowledge resources to succeed for the current pathway. In this way, the linguistic coverage of each of the levels of representation would define exactly how input sentences would be routed through the system.

If the system were to use a more complex pathway than was necessary during translation, then, if properly implemented, output would be produced that could have been created using a simpler pathway. This is an acceptable, if less efficient, result. If the system lacked the linguistic resources to translate a sentence using the sophisticated level of processing it required, then the output would be more English-like in structure than it should. Because most Deaf users of the system would have had experience interacting with hearing people who used non-fluent English-like signing or manually signed forms of English, like Signed Exact English or Sign Supported English, then they may still find this overly English-like translation useful.

5 A Spatial Interlingua for ASL MT

When ASL signers describe a spatially complex 3D scene using classifier predicates, they visualize the elements of the scene as occupying an area of space that is generally within arm's reach in front of their torso. So, signers have a spatial model of the scene under discussion that they can consider when selecting and generating classifier predicates to convey information. An automated system for creating classifier predicates may be able to use an analogous representation.

One way to produce this model is to incorporate virtual reality 3D scene representation software into the MT system's interlingual pathway. After analyzing the English text, the movements of entities under discussion could be identified, and a 3D virtual reality model of the scene could be constructed and/or modified to reflect the information in the English text. This spatial model could serve as the basis for generating the 3D and spatially analogous (topological) motions of the signing character's hands while performing classifier predicates.

Fortunately, a system for producing a changing 3D model of a scene from an English text has been built: the Natural Language Instructions for Dynamically Altering Agent Behaviors system (Bindiganavale et al., 2000; Badler et al., 2000) (herein, "NLI"). The system displays a 3D virtual reality scene and accepts English input text containing instructions for the characters and objects in the scene to follow. It updates the animation so that objects obey the English commands. NLI has been used in military training and equipment repair domains and can be extended by augmenting its library of Parameterized Action Representations (PARs), to cover additional domains of English input texts.

PARs are feature/value structures stored as a library of templates with slots specifying: the agent moving, the path/manner or translational/rotational nature of the motion, terminating conditions, speed/timing, and other motion information. English lexicalized syntactic structures are associated with PARs so that the analysis of a text can be used to select a PAR template and fill its slots. PARs serve as 3D motion primitives and are used as hierarchical planning operators to produce a detailed animation specification; so, they contain fields like preconditions and sub-actions used in NLI's animation planning process (Badler et al., 2000). A PAR generally corresponds to an English motion verb (or a set of related verbs); so, to extend NLI for use in an ASL context, additional PARs will be developed for English motion verbs that often produce classifier predicates.

The MT system's interlingual pathway will use the NLI software to analyze the English source text as if it were commands for the entities mentioned in the text. The NLI can create and maintain a 3D model of the location and motion of these entities. The MT system, unlike other applications of the NLI software, does not care about the exact shape or appearance of the objects being modeled (generic box-like shapes could be used for each). Instead, the location and motion paths of these objects in a generic 3D space are important, since these are used to build classifier predicates.

The MT system would use the spatial model to instantiate a transparent miniature animation of these objects; this animation would be overlaid on an area of the virtual reality space in front of the torso of the character performing the ASL animation output. In the "bumpy road" example, a small invisible object would be positioned in space in front of the chest of the signing character to represent the cat. Next, a 3D animation path and location for the car (relative to the cat) would be chosen in front of the character's chest.

When objects in this "invisible world" are moved or reoriented to reflect information in the English text, the

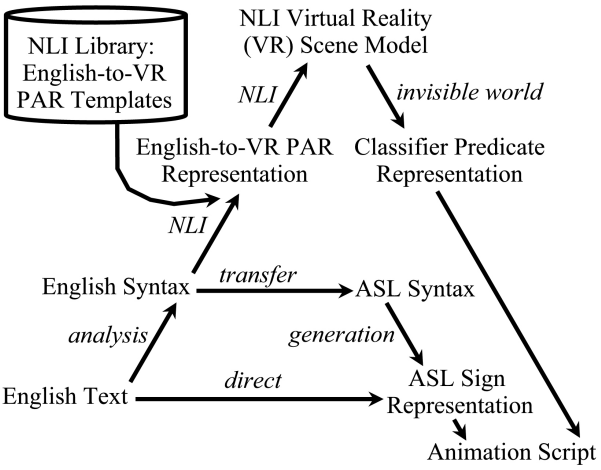

Figure 2: Multi-Path "Pyramidal" MT Architecture.

animated ASL-signing character can position its hand inside of the transparent (possibly moving) object to indicate its new location, orientation, or movement path. By choosing an appropriate handshape for the character, a classifier predicate is thus produced that conveys the spatial information from the English text. Extensions of this design for more complex classifier predicate constructions are discussed in (Huenerfauth, 2004).

This interlingual pathway design would pass along most of the spatial modeling and reasoning burdens to the NLI software, which was designed for this task. It can select relative locations and motion paths for objects in the 3D scene based on prepositions and adverbials in the English input text. It uses collision avoidance, physical constraints, generic and specialized motion primitives, and hierarchical motion planning operators to produce the necessary detail for a 3D animation from the limited information in a corresponding English text.

The full architectural diagram is shown in Figure 2. This design visually resembles the pyramid in Figure 1: direct pathway at the bottom, transfer across the middle, and interlingual pathway over the top of the pyramid. The three paths no longer represent the design choices possible for different systems; they are now processing pathways within a single "pyramidal" architecture.

6 Virtual Reality as Interlingua

The 3D model produced by the NLI software serves as an intermediary between the English text analysis and the classifier predicate generation in this architecture, but that does not necessarily make it an interlingua. In fact, the design differs from interlingual representations elsewhere in the MT literature significantly. To explore this issue, consider a general definition of an interlingua as: a typically language-neutral semantic representation useful for MT that may incorporate knowledge sources beyond the basic semantics of the input text.

First, the model represents those aspects of the input text's meaning significant for translation to classifier predicates; thus it serves as a semantic representation within the 3D motion domain – albeit a non-traditional one due to the ontological simplicity of this domain. Second, this proposed architectural design has illustrated how this 3D scene representation is useful for MT. Third, the NLI software's ability to incorporate physical constraints, collision detection, and spatial reasoning shows how the 3D model can use knowledge sources beyond the original text during translation.

So, the final determinant of this model's interlingual status is its language-neutrality. The 3D coordinates of objects in a virtual reality model are certainly language-neutral. However, ASL linguists have identified discourse and other factors beyond the 3D scene model that can affect how classifier predicates are generated (Liddell, 2003). If the classifier predicate generator needs these features, then the degree to which they are modeled in a language-neutral manner will affect whether the pathway is truly interlingual. Until the final implementation of the generator is decided, it is an open issue as to whether this pathway is an interlingua or simply a spatially rich semantic transfer design. [3]

7 Discussion and Future Work

While English-to-ASL MT motivated the multi-path pyramidal architecture, the design is also useful for other language pairs. Merging multiple MT approaches in one system alleviates the traditional trade-off between divergence-handling power and domain specificity, thus making resource-intensive approaches (e.g. interlingual) practical for applications that require broad linguistic coverage. This architecture is useful when a system must translate a variety of texts but perform deeper processing on texts within particular important or complex domains. It is also useful when the input is usually (but not always) inside a particular sublanguage. Transfer or interlingual resources can be developed for the domains of interest, and resource-lighter (broader coverage) pathways can handle the rest.

While the English-to-ASL system had no statistical pathways, nothing prevents their use in a multi-path pyramidal architecture. Statistical approaches could be used to develop a direct pathway, and hand-built analysis and transfer rules for a subset of the source language could create a transfer pathway. A developer could thus use a stochastic approach for most inputs but manually override the MT process for certain texts (that

[3] Kipper and Palmer (2000) examined PARs as an interlingua for translation of motion verbs between verb-frame and satellite-frame languages. Unlike this system, they did not use PARs within a 3D scene animation; the PAR itself was their interlingua, not the 3D scene.

are important or whose translation is well understood). Likewise, a transfer pathway may use statistically induced transfer rules and parsers, and an interlingual pathway may be manually built for specific domains.

While the pyramidal architecture has applications across many languages, the 3D scene modeling software has benefits specific to ASL processing. Beyond its use in classifier predicate generation, the 3D model allows this system to address ASL phenomena that most MT architectures cannot. The non-topological use of the signing space to store positioned objects or "tokens" (Liddell, 2003) for pronominal reference to entities in the discourse can easily be implemented in this system by taking advantage of the invisible overlaid 3D scene. The layout, management, and manipulation of these "tokens" is a non-trivial problem, and the richness of the virtual reality spatial model can facilitate their handling.

The NLI software makes use of sophisticated human characters that can be part of the scenes being controlled by the English text. These virtual humans possess skills that would make them excellent ASL signers for this project: they can gaze in specific directions, make facial expressions useful for ASL output, and point at objects or move their hand to locations in 3D space in a fluid and anatomically natural manner (Badler et al., 2000). If one of these virtual humans served as the signing character, as one did for (Zhao et al., 2000), then the same graphics software would control both the invisible world model and the ASL-signing character, thus simplifying the implementation of the MT system.

Currently, this project is finishing the specification of the multi-path design and investigating the following issues: deep generation techniques for creating multiple interrelated classifier predicates, surface generation of individual classifier predicates from compositional rules or parameterized templates, and ASL morphological and syntactic representations for the transfer pathway. Another important issue being examined is how to evaluate the ASL animation output of an MT system – in particular one that produces classifier predicates.

Acknowledgements

I would like to thank my advisors Mitch Marcus and Martha Palmer for their guidance, discussion, and revisions during the preparation of this work.

References

R. Bindiganavale, W. Schuler, J. Allbeck, N. Badler, A. Joshi, and M. Palmer. 2000. "Dynamically Altering Agent Behaviors Using Natural Language Instructions." 4th International Conference on Autonomous Agents.

N. Badler, R. Bindiganavale, J. Allbeck, W. Schuler, L. Zhao, S. Lee, H. Shin, and M. Palmer. 2000. "Parameterized Action Representation and Natural Language Instructions for Dynamic Behavior Modification of Embodied Agents." AAAI Spring Symposium.

B. Dorr, P. Jordan, and J. Benoit. 1998. "A Survey of Current Paradigms in Machine Translation." Technical Report LAMP-TR-027, Language and Media Processing Lab, University of Maryland.

J. Holt. 1991. Demographic, Stanford Achievement Test - 8th Edition for Deaf and Hard of Hearing Students: Reading Comprehension Subgroup Results.

M. Huenerfauth. 2003. "Survey and Critique of American Sign Language Natural Language Generation and Machine Translation Systems." Technical Report MS-CIS-03-32, Computer and Information Science, University of Pennsylvania

M. Huenerfauth. 2004. "Spatial Representation of Classifier Predicates for Machine Translation into American Sign Language." In Proceedings of the Workshop on the Representation and Processing of Signed Languages, 4th International Conference on Language Resources and Evaluation (LREC 2004).

K. Kipper and M. Palmer. 2000. "Representation of Actions as an Interlingua." In Proceedings of the 3rd Workshop on Applied Interlinguas, ANLP-NAACL.

S. Liddell. 2003. *Grammar, Gesture, and Meaning in American Sign Language.* UK: Cambridge U. Press.

J. Morford and J. MacFarlane. "Frequency Characteristics of ASL." Sign Language Studies, 3:2.

C. Neidle. 2000. "SignStream™: A Database Tool for Research on Visual-Gestural Language." American Sign Language Linguistic Research Project, Report Number 10, Boston University, Boston, MA, 2000.

C. Neidle, J. Kegl, D. MacLaughlin, B. Bahan, and R. G. Lee. 2000. *The Syntax of American Sign Language: Functional Categories and Hierarchical Structure.* Cambridge, MA: The MIT Press.

É. Sáfár and I. Marshall. 2001. "The architecture of an English-text-to-Sign-Languages Translation System." In G. Angelova, ed., *Recent Advances in Natural Language Processing*. Tzigov Chark, Bulgaria.

d'A. Speers. 2001. Representation of ASL for Machine Translation. Ph.D. Diss., Linguistics, Georgetown U.

C. Wideman & M. Sims. 1998. "Signing Avatars." Technology & Persons with Disabilities Conference.

L. Zhao, K. Kipper, W. Schuler, C. Vogler, N. Badler, and M. Palmer. 2000. "A Machine Translation System from English to American Sign Language." Association for Machine Translation in the Americas.