BUILDING MACHINE TRANSLATION ON A FIRM FOUNDATION

Professor Alan K. Melby, Brigham Young University at Provo, USA

SYNOPSIS

How can we build the next generation of machine translation systems on a firm foundation? We should build on the current generation of systems by incorporating proven technology. That is, we should emphasize indicative-quality translation where appropriate and high-quality controlled-language translation where appropriate, leaving other kinds of translation to humans. This paper will suggest a theoretical framework for discussing text types and make five practical proposals to machine translation vendors for enhancing current machine translation systems. Some of these enhancements will also benefit human translators who are using translation technology.

THEORETICAL FRAMEWORK

How can we build the next generation of machine translation systems on a firm foundation? Unless astounding breakthroughs in computational linguistics appear on the horizon, the next generation of machine translation systems is not likely to replace all human translators or even reduce the current level of need for human translators. We should build the next generation of systems on the current generation, looking for ways to further help both human and machine translation benefit from technology that has been shown to work. Currently understood technology has not yet been fully implemented in machine translation and can provide a firm foundation for further development of existing systems and implementation of new systems. Before making five practical proposals and projecting their potential benefits, I will sketch a theoretical framework for the rest of the paper.

In modern linguistics the following three aspects of language are considered to be very important: syntax, semantics, and pragmatics.

Syntax is the relationships among the elements of a sentence. Typically, a sentence such as "The frog ate three flies" is broken down into a noun phrase ("the frog") and a verb phrase ("ate three flies"). The noun phrase would then be further broken down into a determiner ("the") and a noun ("frog"). However, so far as syntax is concerned, the frog could have been a cat or a rhinoceros or any other noun that is in the same grammatical category as "frog".

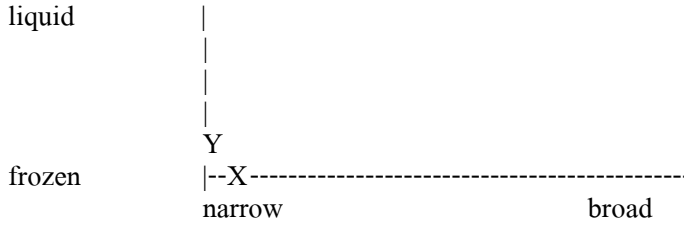
Semantics is that aspect of meaning that can be figured out by looking at just the words of a sentence and their syntax. Each sentence is studied in isolation, without regard for what sentence comes before or after, without taking into account who produced the sentence or why, and without using any world knowledge other than what is encoded in the lexicon as semantic features. Semantics can tell us that the flies got eaten and that the frog was doing the eating but not why someone wants us to know about this event.

Pragmatics is those aspects of meaning that go beyond syntax and semantics. The sentence is viewed as a part of a narration or conversation or other act of communication between humans. We might learn that a child's pet frog had been very sick and was starving but that it now had started eating normally again. Or, alternatively, we may find out that a little boy is telling his father an unconvincing story about what supposedly happened to three

carefully hand-tied items of fishing bait.

Clearly, machine translation must take into account both syntax and semantics. Even in the simplest approach, words in the source language are mapped to words in the target language according to their semantics, and adjustments in word order are made on the basis of syntax. Taking pragmatics into account is a much more delicate topic. In mainstream linguistics, syntax and semantics come first and are separated from pragmatics. A sentence must first be syntactically valid and semantically consistent internally. Pragmatics only selects among options previously approved by the syntactic and semantic components. We will see that sequential application of syntax/semantics followed by pragmatics is not always sufficient to determine whether or not a sentence is valid.

In conjunction with syntax, semantics, and pragmatics, it is very important to consider text type. Typically, we speak of a one-dimensional continuum from technical texts to general texts. See, for example, Snell-Hornby (1), page 32, for a typical one-dimensional classification of text types that has been very useful, ranging from literature and general texts on the left to very narrow domain-specific texts on the right. Here I would like to propose an extension of this approach to text type. Instead of using a single dimension, the proposed classification system is based on two independent dimensions of language. I will call them the specificity dimension and the predictability dimension. If we associate the specificity dimension with the X axis and the predictability dimension with the Y axis, we get the following two-dimensional coordinate system:



Along the X axis, texts at the leftmost extreme are restricted to a single, narrow domain of knowledge. They are specific to that domain. As we move toward the right on the X axis, we find texts that deal with multiple narrow domains and then broader domains until we reach general vocabulary at the rightmost extreme. General texts are not specific to a single, narrow domain. The X axis corresponds roughly to the Snell-Hornby classification system, but reversed, as suggested by my colleague Karin Spalink.

Clearly, every text will use function words such as prepositions and conjunctions. Function words are not specific to a single domain, so if we included function words, no text would be considered narrow. The X axis is concerned with the content words of a text rather than its syntax. If the content words of a text all refer to the concepts of a single, well-defined domain of knowledge, then the text would be assigned a very low X value. If several domains are referred to in a single text or if words of general vocabulary with multiple non-specialized senses are used, then the text would be assigned a higher X value. For the moment, we can distinguish three points, "narrow", "broad", and "in-between", on the X axis. Considerable research will be needed to define a more refined metric which would allow five or ten points along the X axis.

Along the Y axis, texts at the bottom, that is, with a low Y value, whether broad are narrow, are composed of sentences whose meaning can be predicted using only their syntax and semantics. There are no meanings whose selection depends on the situation, such as the meaning of "fly" as a human-created piece of bait rather than a living insect in "The frog ate three flies". There are no unusual syntactic constructions. Everything follows rules precisely.

Low on the Y axis, language is frozen into words and terms with predictable meanings that can be combined in predictable ways, one sentence at a time. As we move up along the Y axis, pragmatics becomes more important in sense selection. Even higher, we encounter more and more creative uses of language. Language becomes more liquid rather than frozen.

Even though the meanings in texts high on the Y axis are unpredictable, they are not random, since they are always seen as motivated when viewed with hindsight. George Lakoff s work on experiential linguistics (2) first brought home to me the important distinction between random and unpredictable. I have since then speculated on the source of this unpredictability (3), but that topic is beyond the scope of this paper. At the higher levels along the Y axis, we see dynamic metaphors and nuances of meaning that are created for the purposes of a single text or conversation and therefore appear in no dictionary. For example, after confessing to his father that he actually lost his father's hand-tied flies and receiving appropriate punishment, the father and son may have made a joke of it and started calling things that disappear mysteriously "frog food". High on the Y axis, pragmatics must be taken into account in parallel with syntax and semantics and not left out or treated afterwards. When a new meaning is created for an expression consisting of two existing words, that new meaning is not as first in the dictionary and not available to the syntax and semantics. If a misplaced pencil were called frog food, a sophisticated semantic rule system would reject the sentence because a pencil is not classified as a food item. But a human could figure it out from the situation or at least know to ask what was meant rather than assuming the reference was an error. No one can predict all future meanings that will be created or which ones will catch on, but, in hindsight, each new meaning is motivated. Suppose the "frog food" example caught on and it became a single word "frogfood". People would start referring to household items that are misplaced as frogfood. This is perhaps unlikely but not much stranger than a child explaining to a teacher that a homework assignment was eaten by the family dog. Didn't the assignment become dogfood?

For the moment, we can distinguish three points along the Y axis: "predictable", "ambiguous", and "creative". Predictable texts would consist of sentences that each have one meaning that can be constructed by combining the elements of the sentence from the bottom up. Any ambiguities are resolved by sentence-internal constraints of the syntax and semantics. Ambiguous texts would include sentences that have two or more meanings so far as the syntax and semantics can tell and require the application of pragmatics. Semantics can resolve many ambiguities by knowing which domain or domains a text refers to. If a text discusses chemistry but not the sport of baseball, then a base would probably be the opposite of an acid rather than one of the four points of a baseball diamond. But a text describing the preparation of fish for eating could refer to a scale in a way that would be ambiguous between a device for weighing and the thin plates that form the outer layer of the body of a fish. Both meanings of "scale" are in the dictionary, but pragmatic context is required to choose among them, so this ambiguity would push the text containing it a little bit up on the Y axis. A text even higher on the Y axis, a creative text, is one that contains dynamic metaphors, analogies, syntactic constructions, or other non-standard aspects that suggest meanings that would not be predicted by the syntactic and semantic components. Here it is not just a case of pragmatics choosing among multiple computed meanings. There are meanings that cannot be computed; they are created on the spot, so they cannot be presented to the pragmatic component by the syntax and semantics. Such meanings are highly unpredictable, even though they make sense to a human. Much research is needed to identify how often unpredictable meanings occur in various texts and to decide how to assign a value on the Y axis to a particular text. What if parts of a text are very predictable and other parts are unpredictable? Each unpredictable meaning of a word, expression, or construction is a barrier to predicting the meaning of the entire text. How does one assign a composite Y value to an entire text? For the moment, we can just use the three values of predictable, ambiguous, and creative.

Even though additional research is needed to elaborate this two-dimensional view of the properties of texts, we can already use the notions of specificity and predictability to discuss machine translation. Machine translation has always done best on narrow texts, restricted to a single domain of knowledge, with predictable meaning and rigorously controlled use of syntax and vocabulary. This type of text is sometimes called a controlled domain-specific text. In the framework just described, controlled domain-specific texts would be in a zone located in the lower left-hand corner near the origin of the space of text types defined by the two axes. Dynamic general language would be higher up than controlled language and toward the right. I have written elsewhere on the contrast between controlled domain-specific language and dynamic general language. See Melby and Warner (4).

In the past, I have emphasized these two extremes, controlled domain-specific and dynamic general, and I have claimed that there is a wall between them. Others have suggested, rightly, that there is a continuum of text types rather than just two sharply contrasted types, which would suggest that there is no wall. The two dimensional approach allows the "wall" to be identified with the Y axis. On a two-dimensional scale, two points can be close together on the X axis yet far apart on the Y axis. This Y-distance, the combined effect of many individual barriers to predictable meaning, could be metaphorically referred to as a "wall", but it depends on X and Y being independent dimensions. If specificity (the X axis) and predictability (the Y axis) are independent dimensions, then it should be possible to find examples of texts that fit into all four corners. If, on the other hand, only two of the four corners get filled (lower left and upper right) then the dimensions are not independent and we should go back to the traditional one-dimensional model for text types.

Although much more research is needed, we can already imagine texts that would fit into the upper left-hand corner. They would be texts that explore new domains, creating terminology as they go, or presentations of a technical topic for a general audience by a skilled and entertaining popular science writer. The lower right-hand corner would be the unfortunately voluminous category of rather boring material that is neither highly specialized nor unpredictable. Such texts may transfer information but lack what an English composition teacher might call "life" or "feeling". In the middle of the space of text types toward the right we find general language texts and toward the left we find what is called LSP (Language for Special Purposes) texts. LSP texts that are assigned further down toward the left are more amenable to high-quality machine translation. If they are further up and to the right, they are likely to be much less amenable. Many LSP texts occupy a gray zone in the middle that makes them more or less amenable. This preliminary exploration fills out the four corners and middle of the space of text types and suggests that specificity and predictability are independent dimensions. A text type occupies a fuzzy region of the space of text types, not just a fuzzy line segment on a single continuum.

The above two-dimensional approach to classifying texts is relevant to machine translation since it provides a richer system than a one-dimensional continuum for explaining why some texts do well when machine-translated and others do poorly. One way to put it is that the further one strays from the origin the more problematical the text is likely to be for machine translation.

PRACTICAL PROPOSALS

Within the above framework, I can make some practical proposals for further development of machine translation on the basis of what has been shown to work rather than exploring dangerous territory.

1. Consumer Labelling for Machine Translation Systems

My first proposal applies both to current and future machine translation systems: machine translation vendors should be open about what kind of text their system is intended to translate and for what purpose. Presently, there are two very different uses for machine translation.

One major use for machine translation is for what I call "indicative translation". Others have called this same use "gisting" or "information gathering". An indicative translation is not intended to be a high-quality translation comparable to the work of a professional human translator. Its purpose is only to provide a rough indication of the content of a document written in a language that the user does not read. An indicative translation into the user's first language may satisfy the needs of the user relative to the document in question, or it may be used as a basis for deciding whether to commission a high-quality human translation of the document.

In terms of the X-Y framework for text types, a machine translation system for indicative translation can take on any kind of text and likely produce a somewhat useful, though rather ugly result. The ugliness usually stems from the way the raw output is produced. The source language is manipulated by replacing source-language words with target-language words. As Humphreys put it in his paper at the 1991 Aslib conference on Translating and the Computer (5), such systems perform "extensive cosmetic surgery" on the source text, rather than really translating it. This description may be overly unkind; nevertheless, the result is a non-natural language which is interpreted by the user as natural language thanks to the intelligence of the user, not the intelligibility of the raw output.

A second major use of machine translation is for what I call "publication-quality translation". A publication-quality translation must be comparable to a professional human translation that is intended for publication. Raw machine-translation is seldom up to publication quality before post-editing by a skilled human bilingual. In using machine translation as step toward publication-quality translation, an important economic and human issue is the amount of post-editing required to bring the raw output up to publication standards. If there is too wide a gap between the raw output and the finished product, the post-editing process will be excessively expensive for the user and cruelly tedious for the human post-editor.

Relative to the X-Y framework, current machine translation systems for publication-quality translation are highly sensitive to text type. Usually, in order to keep post-editing down to a reasonable level, the source text must be an example of controlled language in the lower right-hand zone, such as a Xerox photocopier maintenance manual. If the source text is more general (toward the origin, that is the left, on the X axis), then it must be very low on the Y axis. More likely, however, a machine-translation system for publication quality will be tailored to one knowledge domain at a time. This tailoring involves access to a domain-specific terminology database.

A combination of indicative and publication-quality translation is conceivable. Some

human post-editors, but not all, are able to make limited corrections to an indicative translation, keeping the cost of post-editing down while making the text somewhat more readable.

Just as food is labelled in some countries to indicate nutritional content as a service to the consumer, perhaps a machine translation system should be clearly labelled as to the type or types of text it is intended to translate and whether it should be expected to produce indicative translation or provide a basis for publication-quality translation and what amount of post-editing should be expected. Of course, no system can produce useful output if the dictionary is not well-made and appropriate. But it is unlikely that a single machine translation system would be equally suitable for both indicative and publication-quality translation. A publication-quality system would be lost in the variety of texts presented to an indicative translation system, and an indicative translation system would not be able to produce sufficiently high quality raw output because it would not take advantage of the properties of a controlled-language text. My first proposal is simply that vendors openly admit that a given system is not equally suitable for all tasks and that systems be labelled accordingly.

2. A Formal Definition of Grammar for Controlled Languages

Consider a machine translation system that is labelled for use in producing publication-quality translations of controlled-language texts. Such systems usually attempt a complete syntactic analysis of each sentence of source text. As mentioned above, the successful use of such a system requires minimizing the amount of post-editing required on the raw output. One factor that heavily influences the ability of the system to perform syntactic analysis, a factor that directly affects the quality of the raw output and thus the amount of post-editing, is the degree of match between the syntactic structure of the source sentences and the formal grammar that defines what structures the system is expecting. Note that the system is not expected to accept all sentences of the source language that a human would accept. By definition, a controlled language is a formal language. Syntactically that formal language resembles naturally produced human language but is not identical to it.

My second proposal is that computational linguists create an explicit formal definition of syntax for controlled languages that can be studied independent of any particular computer implementation. Such a definition could be as simple as a context-free grammar such as those used in computer science to define the syntax of programming languages. Another possibility would be to use some branch of mainstream Generative Grammar, such as the Principles and Parameters approach or the Head-driven Phrase Structure approach, or a branch of Dependency Grammar.

Along with the core definition would be a mechanism for modifying the definition to allow or disallow certain constructions and for testing the modified grammar for proper format and internal consistency. Essential to this already substantial task is the non-trivial task of developing verification software that can be used to check a source text for compliance with a formal definition. Some computer programming languages have such verifiers, sometimes called "lint" programs. This controlled-language syntax checker should be used as early on in the document production chain as possible, preferably by the author. In some environments, the syntax checker could even be integrated with the word processor that is used by the author to create the text. What is called a grammar and style checker in today's word processors does not go far enough toward a complete syntactic analysis, but is certainly an important step along the way.

The grammar of a controlled language should be carefully crafted to strike a balance that reduces ambiguity without so limiting the range of allowable constructions that authors feel suffocated.

Some syntax checkers do exist, but they are still too proprietary. I am calling for a public standard for defining the grammar of a controlled language. The same grammar would be used by both a syntax checker and a controlled-language machine translation system. This is a big project, but an appropriate one for a university setting, where it could be viewed as a public service.

If text is re-checked for compliance to the syntax expected by the machine translation system just before it goes into the system, then there should be no syntactic analysis errors, and the quality of the raw output should be noticeably higher than if the syntax of the source text is not pre-checked. One does not expect top performance from an automobile if the type of fuel does not match the design of the engine. Diesel engines should be fed only diesel fuel.

It is also possible to think about grammar checkers that are told which domain or domains are allowed in the source text and thus enter into the territory of semantics, detecting ambiguities of reference caused by allowing multiple domains. As a colleague, Klaus Schubert, has suggested, an ideal grammar checker for machine translation is a filter that tells you whether the text is likely to produce high-quality raw machine translation and, if not, how to remedy that.

3. Format Use During Translation, Not Just Format Preservation

Typically, the format of a translation is expected to match the format of the source text. It is well-established that in this case a machine translation is more effective if it can preserve the format of the source text in the raw translation.

My third proposal for improving machine translation systems is a public standard to facilitate not the preservation of formatting information during machine translation but the use of that information to improve the quality of the raw machine translation. The obvious basis for such a standard would be SGML. SGML is an international standard for formally defining the structure of a class of documents. Specifically, my proposal is to develop a method of associating elements of an SGML DTD (Document Type Definition) with an inventory of element types (such as headings and bibliographic references) that are useful to a machine translation algorithm. One of the best-known applications of SGML is called HTML and is used to define the structure of a page on the World Wide Web.

A simple example of an element type would be a heading. Headings are often noun phrases rather than full sentences, and the machine translation grammar can benefit from knowing that a particular piece of text is a heading. Another simple example is to distinguish among various uses of quotation marks. A literal quotation of what someone said should be treated differently from a quotation that indicates an unusual or made-up word. Using SGML, the above distinctions can be made available to guide computer processing even though not all the distinctions are visible in the presentation of the text to the end user. Some work on the use of format during machine translation is being carried out at Carnegie Mellon University. Hopefully, some of the fruit of such projects will soon become available to all machine-translation developers.

To preserve the format of a text in the output text without taking advantage of it during the process of translation, is, from the perspective of the machine translation system, like eating the skin of a peach and throwing away the inside.

Admittedly, it may be a long time before all source texts are marked up in SGML when they are authored, but there is a portion of this proposal that could be implemented very soon. In addition to marking up the body of a text with format codes, there should be a standard translation request header that comes before the body of the text and that would include the language of the source text, the desired target language or languages, and other specifications of how the translation should be processed by a machine (or by a human, for that matter).

4. Universal Terminology Interchange

No matter how well the syntax of a controlled-language source text has been checked and its format specified, the raw translation cannot be of high quality unless an appropriate bilingual terminology database (sometimes called a "termbase") is available to the machine translation system. An appropriate termbase is one that contains source-language term pairs that match the domain of the source text. The quality of the termbase may well be more important to the quality of the raw output than the quality of the syntactic analysis component. A present hindrance to the use of terminology is the diversity of formats in which termbases are laid out. This diversity makes it difficult to use a termbase with several different machine-translation systems.

My fourth proposal is the development of a standard format for the interchange of terminology in machine-readable form.

I am part of an international effort to define a terminology interchange format that could be used by both computer tools for human translators and machine translation systems. This effort, which addresses my fourth proposal, is based on two other projects: MARTIF (a project of Technical Committee 37 of ISO) and TRANSTERM (a European Union project). An entire paper could be written on this topic. Indeed, a workshop on this topic is scheduled for the International LSP Conference next August in Copenhagen.

A widely-accepted standard format for terminology interchange would allow the transmission of terminology along with a source text. In most source texts, some specialized terms need to be translated consistently. Concept entries for those terms would be extracted from an appropriate termbase and passed along with the document in a universal interchange format. That termbase subset would then be converted, sometimes automatically, to the internal representation used by the machine-translation system and consulted during the machine-translation process. Of course, we are not talking about function words (as mentioned previously, function words are grammatical markers such as prepositions and conjunctions). Function words are coded very differently in different machine-translation systems and it would be very difficult to find a universal format for the information needed about such words by machine translation systems. Initially, verbs, adjectives, and adverbs would not be interchanged either. However, specialized terms are typically nouns or phrases that can be treated as if they were a noun. Nouns are more straightforward to encode for use by multiple systems. Even so, the universal encoding of the various features used by machine translation systems even on nouns would be an extremely challenging project. It remains to be seen whether a terminology interchange format will permit automatic interchange among different machine translation systems, but the rewards justify a considerable effort in this regard and even a partially automatic interchange might be worthwhile. An interchange format would also allow shared development of a very friendly dictionary update module.

5. Links Between a Source Text and a Termbase

Once most machine translation systems are taking advantage of the format codes in the source text and most source texts are delivered with machine-readable termbase subsets, there is one more step that can be foreseen.

My fifth proposal would be to mark terms in the source text at authoring time or soon thereafter with the help of an editorial assistant and link them to a termbase. This would improve the quality of the source text and facilitate later translation. Using SGML, a term can be marked unambiguously in the source text, even if it consists of several words, and the markup need not be visible in the presentation of the text to the end user.

POTENTIAL BENEFITS

The five proposals just made are ambitious yet realistic. They assume no breakthroughs in linguistic theory or computer software or hardware. They will require a lot of hard work and a spirit of cooperation among developers, translators, and linguists. However, the potential benefits are substantial.

Consumer labelling for machine translation systems would reduce disappointment by users. Systems for gisting and systems for producing publication quality output are not interchangeable. A standard way of defining syntax for controlled languages would permit higher quality output. And it would permit easier comparisons of two systems if they both accepted the same formal grammar. For publication-quality systems, the emphasis should not be on attempting to continuously broaden the range of structures that are accepted. Neither should next-generation production systems attempt to produce high-quality output for texts that are high on the Y axis of our two-dimensional scheme. The emphasis should be on defining controlled languages that are as restricted and easily processed as possible while allowing sufficient structures and concepts to express what needs to be said. Then effort can be put into improving the quality of the raw output for a given controlled language.

The third, fourth, and fifth proposals (marking and using SGML format codes in the source text, accompanying the source text with a termbase subset in a universal format, and linking the text with the termbase subset) would benefit not just machine translation and post-editors but also human translators who do not use machine translation. These last three proposals would allow more sophisticated and effective translator productivity tools.

Much is said these days of the impact of the Internet on translation. An electronic package consisting of a source text with a translation-specification header and a termbase subset would be ideal for submission to an indicative machine translation system somewhere on the Internet or to a free-lance translator anywhere in the world.

REFERENCES

Thanks are expressed to the colleagues who provided valuable feedback on this paper, especially Roald Skarsten, Karin Spalink, Klaus Schubert, Michael Sneddon, and Arle Lommel. The mixture of American and British punctuation was my choice, not their suggestion.

1. Snell-Hornby, Mary. 1988. Translation Studies: An Integrated Approach. Amsterdam/Philadelphia: John Benjamins Publishing Company.

2. Lakoff, George. 1987. Women, fire, and dangerous things: what categories reveal about the mind. Chicago: University of Chicago Press.

3. Melby, Alan and C. Terry Warner. 1996. Translation and Free Will. A paper presented at an international symposium on historical and theoretical aspects of translation held at the Geneva school for translators and interpreters in October 1996.

4. Melby, Alan K. with C. Terry Warner (1995) The Possibility of Language. Amsterdam: John Benjamins Publishing Company.

5. Humphreys, R Lee. 1992. Proceedings of the 1991 Aslib conference "Translating and the Computer 13", page 93. London: Aslib.