Do What I Mean: Issues in Cross-lingual Collaboration

Florence Reeder
MITRE Corporation / George Mason University
1820 Dolley Madison Blvd.
McLean, VA 22102
freeder@mitre.org[1]

Keywords: natural language processing, cognitive models, collaboration

Abstract
In most virtual worlds and virtual communities, language is at the heart of communication. When we extend these communities to the international stage, we are faced with challenges in interaction. We will not examine the infrastructure supports beneath multi- and cross-lingual communication. Instead, we look at the next level of enabling communication – that of automated machine translation (MT) of the utterances. MT is a complex and, it will be argued here, an endogenous problem. Even though the MT process is endogenous, we are not prevented from creating useful interfaces and supports for cross language interaction in virtual worlds and simulations. This paper starts with a basic description of machine translation, the problems of MT development and why MT can be categorized as endogenous. Afterwards, we describe the classes of problems that MT must solve which are specific to virtual words. Then, we looks deeply into the communicative system and start to describe a framework to handle one particularly pesky problem of MT: the process of building language-independent representations to enable effective translation.

INTRODUCTION

In most virtual worlds and virtual communities, language is at the heart of communication. Yet this new field of linguistics has not been greatly studied – from a practical, a sociological or a computational point of view. It represents a unique form of communication with characteristics of both spoken and written text, including the extensive use of acronyms (such as ROTFL), jargon, specialized terminology, and specialized syntax. When considering virtual worlds/communities in a multilingual world, it therefore presents a new set of challenges, since not all participants will be capable of communicating in the same language.

While there are many nuts and bolts issues to multilingual and cross-lingual interaction in virtual worlds, such as keyboarding, display and language representation, these have been discussed elsewhere [e.g., Reeder & Harper, 2000] and will not be addressed here. Instead, we look at the next level of enabling communication – that of automated machine translation (MT) of the utterances. MT is a complex and, it will be argued here, an endogenous problem. Even though the MT process is endogenous, we are not prevented from creating useful interfaces and supports for cross language interaction in virtual worlds and simulations. This paper starts with a basic description of machine translation, the problems of MT development and why MT can be categorized as endogenous. Afterwards, we describe the classes of problems that MT must solve which are specific to virtual words. Then, we looks deeply into the communicative system and start to describe a framework to handle one particularly pesky problem of MT: the process of building language-independent representations to enable effective translation.

MACHINE TRANSLATION

Machine translation systems translate between human, natural languages. This section describes Machine Translation (MT) by first showing its origin. A brief depiction of the processing involved is followed by a demonstration of the difficulties in translation which help to characterize it as a bizarre process. Weaver [1947] describes the translation process as a kind of noisy channel decryption process: the French document is really an English document coded in French.[1] For instance, the literal translation of **omelette du fromage** is **omelette with cheese**. So far, the process is simple enough. Yet, as anyone fluent in more than one language knows, the process of translating between languages is more than the substitution of one word for another. Consider the example **Parlez vous Français?** Literally translated, it is **Speak you-formal French? –** understandable, but not quite good English. More analysis of the source or better generation of the target is necessary for quality translation.

Traditionally, translation systems have been composed of several components or processing layers. The process is frequently described in terms of a pyramid [Vauquois, 1968], **Fig. 1**. To show the processing, a system begins at the left side and analyzes the source language – the analysis is more complex as we ascend the pyramid. At the transfer point in any given system, it changes the results of the analysis into a format suitable for the generation of the

[1] "When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.' " [W. Weaver, March 1947]

target language. The system then descends the pyramid into the target language. Interlingua represents a pure translation process where analysis yields a language-independent representation from which generation can be done. The components which address these different levels include: a bilingual lexicon (or list of words), a grammar of the source and target languages, a set of transfer rules which translate structures between the source and target representations.

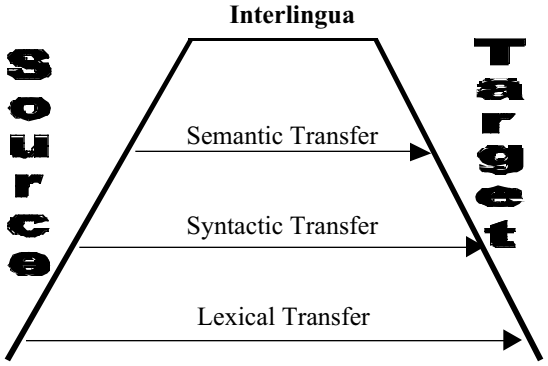

Figure 1 The Machine Translation Pyramid

To demonstrate the difficulties inherent in translation and why it resists traditional processing, consider the word **bank**. It can mean a financial institution, a side of a river, the act of counting on an event, or driving around a corner. Translating this word into foreign languages is not straightforward, as other languages have different words corresponding to the different meanings of **bank**. Because of this many-to-many mapping of words, word for word substitution is insufficient, therefore, a lexical transfer system translates poorly.

The next level of representation, the syntactic level, provides some disambiguation. For instance, determining if **bank** is a noun or verb contributes sufficient evidence to cut the number of readings in half. The grammatical categorization of words is accomplished through stochastic models of language or through rule-based analysis. The problem of accurate translation remains as we still have multiple meanings for the nominal **bank** and the verb **bank**.

Further up the pyramid, the semantic level, information about possible meanings of words contributes to the translation process. To continue with the banking example, it is possible to differentiate **bank** by looking at neighboring words (a, b). Semantic disambiguation can be addressed through probabilistic translation models utilizing n-grams (word pairs or triplets). Yet this is also not optimal as neighboring words do not necessarily provide sufficient information to assist translation in as in (c). In all three examples (a-c), the words **bank** and **left** occur in the sentence, yet the meanings are very different.

(a) First National Bank is on the left.

(b) The Left Bank is the site of the first national parade.

(c) I left the bank yesterday.

Another way to try to pick the correct word sense is to develop a large ontology of concepts and meanings to support the translation process.

As we have moved up the translation pyramid, the necessary amount of target and source background knowledge has increased as has the preparation cost. Also, we still do not have sufficient representations for the transformation of language at the conceptual level. To say **I am banking on the horse coming in** implies a concept which literal translation may not be able to account for. In this case, building an interlingual representation is an appropriate choice.

INTERLINGUAL REPRESENTATIONS

An interlingua is a "language independent" representation. While it represents long-term cost savings in translation[2] systems, it also demands much more complex up-front knowledge acquisition and engineering. Even more problematic is the notion of what it means to be a true interlingual representation. Consider the following possible interlingual representations shown in **Table 1**.[3] Each of these is an approach to interlingua, but all of them are language-dependent. The degree to which they are language dependent reflects how they fit into their domain.

Kipper and Palmer [2000], for instance, developed an interlingua for providing a multilingual interface to a simulation. Their interlingua reflects the needs of the physical environment and the planning system which operates in it. While reasonably language independent (if one grants that the primitives are language independent), the amount of work necessary for this very limited domain was substantial. Levin et al [2000] measure the notion of language-independence when they test for inter-coder agreement.[4] While they had interlingual coders in different countries, all of them were English speakers and they related difficulty in getting non-English speakers to code the same information accurately. Having described the processing necessary for MT and illustrated the difficulty inherent in it, we will now argue that MT is an endogenous process that requires a new notion of interlingua to effectively reach the next level of capability.

Table 1 : Interlingual Representation Examples

---

[2] If M is the number of source languages and N is the number of target languages, one only needs to develop M+N instead of M*N systems.

[3] Note that each of these supports a different notion or function.

[4] There is a standing joke in the MT community, that an interlingua is just English with upper case letters, e.g., **dog** ⇔ DOG ⇔ **chien**.

```
COMPANY: [ NAME: N [text]
HEADQUARTRS: H [office]
SUBSIDIARIES: S [set: company]
EMPLOYEES: E [set: human]
SALES: V [currency] … ]
```
(Farwell & Helmreich, 2000)

```
A:give-information+price+room
(room-type=double,
       price=(quantity=150,
              currency=dollar,
              per-unit=night))
```
(Levin, et al., 2000)

```
Activity: [ ACTION ]
Participants:[ agent: AGENT] [objects: OBJ1,
OBJ2]
Applic_cond: [reachable (OBJ1)] [have(AGENT,
OBJ2]
Preparatory_spec:  [get (AGENT, OBJ1)]
Termination_cond:  [contact (OBJ1, OBJ2)]
Post_assertions:  [contact (OBJ1, OBJ2)]
Path, duration, motion, force
Manner  :  [ MANNER]
```
(Kipper & Palmer, 2000)

MT AS AN ENDOGENOUS PROCESS

Past approaches to MT system development have relied on traditional knowledge representation ideas, based on notions that language is reducible for processing. Therefore, they break down a paragraph into sentences, sentences into phrases, phrases into syntactic parts, syntactic parts into words and words into meaning. Each of these can be accomplished through a rule-based or statistically-based method which is based on the notion that words are independent. Yet, it is our argument that this is not a realistic view for MT systems. The sentiment is not new in the MT community.[5] We will now show how MT systems rightly fit into the category of endogenous systems, an argument for a new approach to representing information necessary to build them.

MT qualifies as endogenous system [Kercel, 1999] because of the partially entangled behavior of words. For instance, meanings can be compositional: **cheese fries** literally interpreted is either **fries with cheese on them** or **cheese is a substance that can be fried**. At the same time, we have many wordings that are not compositional, defying a logical representation. The saying **He bought the farm** has a non-compositional, or idiomatic, meaning that is different from the sum of the parts. Yet, there is a reasonable and rational causality to idiomatic language usage. The phrase **thumb-rule** stems from the days where the **Rule of Thumb** determined the size of a stick with which a man could beat his wife.[6] So the entanglement of words is one reason for considering MT under the endogenous system model.

If MT can be viewed in the impredicative model, then we can apply aspects of endogenous modeling to improve MT systems – such as a record of the past; prediction of the future; inferential elements; causal elements in our ontology; and anticipatory behavior. A record of the past can be gleaned from corpora as in statistical language models. Ontological representations with causal elements under the bizarre model should contribute sufficient evidence of relationships to support language independence. To achieve the next level of capability, we require a new way of looking at computing translations because we are modeling a process that is impredicative. The question, then, becomes how to utilize the best of the traditional models? Or should we even try this? What are the pieces of this overall puzzle? Can we find partial solutions that are effective and efficient?

We will now look at the specific requirements for MT in virtual worlds, as observed in preliminary development. Afterwards, we present ideas to answer some of these questions through a proposed model of evidential lexicons.

MT IN VIRTUAL WORLDS

From a linguistic point of view, virtual worlds and collaborative environments represent a new and exciting area of research. Linguistic analysis has typically fallen into one of two camps – spoken interaction between two or more active participants or written interaction for passive participation. Each of these has features which delineate it from the others and there is little overlap between the two. Virtual worlds and collaborative environments, however, yield a new form of interaction – one which has some characteristics of spoken interaction and some characteristics of written interaction. This section looks at this emerging area of linguistic analysis and describes the resulting implications for cross-language support in these environments.

The multilingual challenges faced by collaborative environments include this aforementioned unique interaction style. We will shortly look more deeply at this. Also thought-provoking are the kinds of specialized needs of the interactions; the importance of getting the interaction right; the difficulty of capturing and analyzing actual language use; and the inherent difficulties of MT use in any environment. Each of these will now be examined in greater detail.

[5] "… as to the problem of mechanical translation, I frankly am afraid the boundaries of words in different languages are too vague … to make any quasimechanical translation scheme very hopeful." [N. Weiner, April 1947]

[6] No greater than the width of his thumb.

The interaction in a collaborative environment exhibits characteristics of both written and spoken exchanges. For instance, ellipsis is very common in this environment:

"Do you have the AK-47?"
"In my bag"

Since humans will take advantage of the shortcuts provided by both domains, the processing necessary to adequately handle the interaction type is much more demanding. Further analysis will tell us the magnitude of this, but we have already seen ellipsis (the cutting out of part of a sentence as understood), sentence fragments (one word answers to questions), informal language structure and extended negotiation of meaning.[7]

Other specialized language needs are more traditional MT problems. These include acronym use, proper name use, command structures and accelerated jargon development. Acronyms fall into two categories – those more straightforwardly interpreted and those where analysis and user input will have to be included. For instance, **NATO** is translated as **OTAN** in French. This is because the organization is translated, but the acronym is not. Other organizations/acronyms may or may not follow this convention. For instance, what is the proper translation of **ROTFL** (Rolling On The Floor Laughing) for a French Speaker? One system would tell you **RSLPR** (Roulement Sur Le Plancher Riant), yet a French speaker would probably use an entirely different acronym **FCQVDJ** (Faites Ce Que Ceux Dire Je –> "I am doubled over laughing").

In addition to acronyms, proper names represent a complication for translation. The difficulty with collaborative environments may be that the context will be either more complex or lacking entirely. In a test of 100 news articles translated from Spanish to English, 2600 named entities were found. The two human translations available agreed only 90% of the time on the proper translation of names. Both of these factors are compounded by the almost dialect quality of most virtual worlds and environments. This social phenomenon will be discussed shortly, but the direct implication is a need for careful analysis of typical dialogs to ensure that new jargon is captured in the translation system.

Even once the lexical demands are met, we are still faced with structural issues. For instance, distinguishing commands to the system from social interactions (and which needs to be translated) is one discourse aspect of the problem. Some environments, such as military collaborative ones, will necessarily need specialized interaction structures. While this constrains the MT in one way, it implies that the text may not be the clean well-formed sentences expected by MT. Some situations will demand nearly 95% accuracy of translation – an unrealistic expectation for current technical capabilities. Aside from the words, it is also desirable to maintain expressed relationships between entities and to handle social or cultural needs right. There is a great difference between:

Bill likes Mary.
Mary likes Bill.

Maintaining the correct relationships is necessary. Socially, mistranslations can be comical – but also detrimental to building good rapport in the communities as expressed by:

You are invited to take advantage of the chambermaid.

This also can be seen as an idiomatic use of language. Idioms, as described above, are phrases where the sum of individual word meanings differ from the overall meaning of the sentence. In these cases, the translation, while technically correct, fails to capture the essence of the meaning. The resulting sentence "just doesn't sound right."

The best method currently available for MT progress is to capture sample interactions, texts and translations and to use these to update and upgrade the MT system. We will describe a model for how the computation of this may be improved. There are social challenges to incorporating MT into virtual worlds and simulations to this approach. The biggest of these is the social implications of monitoring conversations, recording interactions and using them.

Even clearing this hurdle, the sociological aspects of language use in these environments stand out. Language is tied to our social identity. We change our within-language use depending on our social status in the interactive community. Will providing MT support this or erode it? Also, language is seen as a membership criteria for some groups. That is, specialized jargons develop in areas to distinguish the experts from the novices. To fully qualify for membership, you have to be able to speak the jargon. While MT could enable this (assuming we can get a handle on the unique language use in each community), there may be social repercussions from lowering the bar of admittance. Still, we will look at the potential for good and start describing a possible model for building lexicons which are better able to support MT development.

EVIDENTIAL LEXICONS – A MODEL?

The questions presented earlier aim to find ways to arrive at a model for MT lexicons and computing which reflect the endogenous paradigm. While we are still grounded in computation,[8] we believe that there are reasonable approaches to explore the questions. Part of the answer may be found through evidential reasoning models[9] and psycholinguistic research. The end goal is a representation which supports the combination of multiple

[7] It should be said that this could be an advantage over more passive forms of MT use which tend to present the meaning without benefit of feedback.

[8] Because as engineers, we do have to build something.

[9] Schum [1994] describes many of these.

pieces of evidence that contribute to the "meaning" of a concept in a language independent way.

Because many biological systems are endogenous, we look to psycholinguistics for ideas in modeling the impredicative nature of MT. Psycholinguistic models suggest why we select the words and grammatical constructs we use. Language is one of our primary means of acquiring information – through talk, through reading, through words. Language is not the form of the representation, but words and the combination of these words are part of our knowledge structure. The notion that we remember a gist of what was said in place of the exact words means that there is something else going on. Professional translators show this often when they translate sentences according to general meaning instead of for exact words. In fact, it has been shown that there is a wide degree of variability even in translating "factual" reports [Farwell & Helmreich, 1996]. Another study [Al-Onaizan, et al., 2000] reports that translation can be a process of picking the words you know, guessing a context and applying that context to unseen words. There is reason to suspect that the translation process is different than the sum of the parts.

Connectionists [e.g., Winograd & Flores, 1986; Cummins, 1989] model this associations of words with neural networks. Taking from this tradition, Jurafsky [1996] treats lexical organization with probabilistic or evidential models. Words, and therefore some aspects of knowledge, are organized into associated structures in the mind.[10] The following experiment describes that this association exists and that it is probabilistic in nature:

> "As expected, people pressed the button faster when recognizing *ant*, which is related to *bug,* than when recognizing *sew*, which is unrelated. Surprisingly, people were just as primed to recognize the word *spy*, which is, of course, related to *bug*, but only to the meaning that makes no sense in the context." [Pinker, 1994]

Our recognition and usage are sensitive to stronger associations, so a word is more easily recognized if it is related to a previously stated word. New information is learned by incorporating it into the existing network or strengthening or renewing current connections. This knowledge structure can then be processed via following the links and connections. The links have strengths which indicate their associative force. Koestler [1964] is one of the first to show the biological basis for this, indicating these are useful phenomena to incorporate into a lexical model.

In describing this possible model, we realize that space does not permit us to develop the idea fully. If we assume our lexical model to be evidential with both traditional and bizarre data, the core evidence for a translation of a word comes from dictionaries. These are the evidential equivalent of jurisprudence representing the work of people whose job it is to know how words are used and translated. Yet these dictionaries contain ambiguity in both meaning and usage of words. Therefore, we would augment a dictionary entry with additional information which is learned from multiple evidence sources both traditional and bizarre: associated words (n-gram measures); part of speech measures; syntactic measures; morphological information; ontology information. In selecting a translation, the evidence is combined to find the most likely translation – based on a number of sources.

In a basic representation, the evidence could be combined using Bayesian networks. This follows the Jurafsky model from psycholinguistics. We have not divorced ourselves from the need to use language to label nodes in the network. We are, however, more language neutral than previous models because of the ability to combine evidence sources. For instance, the Japanese distinguish the different forms of rice (cooked, uncooked or in the field). One could easily envision a Japanese lexicon where the form of the rice has a great enough weight to allow for the correct word to be chosen. More advanced models could rely on other evidential models [Schum, 1994], although this is an area for further exploration.

CONCLUSIONS

We have described machine translation as an endogenous system especially in the area of interlingual representation. We have very briefly presented a possible computational solution through evidential reasoning. The coming year(s) will enable us to perform deeper analysis and begin building these lexicons. We would like to thank our reviewers for their astute comments.

[10] Note that Chomsky [1972] is the first to see this relationship of pre-wired language structures and it is from Chomsky that Pinker [1994] derives much of his work.

References

Al-Onaizan, Y., Germann, U., Hermjakob, U., Knight, K., Koehn, P., Marcu, D., Yamada, K. 2000. Translating with Scarce Resources. In *Proceedings American Association for Artificial Intelligence Conference, AAAI-2000*.

Chomsky, N., 1972. *Language and Mind*, Harcourt, Brace and Jovanovich.

Cummins, R., 1989., *Meaning and Mental Representation*, MIT Press.

Farwell, D. and Helmreich, S., 2000, An Interlingual-based Approach to Reference Resolution, In *Proceedings of the Workshop for Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP*, ANLP/NAACL-2000, Seattle Washington.

Helmreich, S. and Farwell, D. 1996. Translation Differences and Pragmatics-Based MT. In *Expanding MT Horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA-96).* Montreal, Canada.

Kercel, S., 1999, Embedded Intelligence: Bizarre Systems for the Steel Industry, Presentation to AISE – Sensors Meeting.

Kipper, K. and Palmer, M., 2000, Representation of Actions as an Interlingua, In *Proceedings of the Workshop for Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP*, ANLP/NAACL-2000, Seattle Washington.

Koestler, A, 1964, *The Act of Creation*, The Penguin Group.

Levin, L., Gates, D., Lavie, A., Pianesi, F., Wallace, D., Watanabe, T. and Woszczyna, M., 2000, Evaluation of a Practical Interlingua for Task-Oriented Dialogue, In *Proceedings of the Workshop for Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP.* ANLP/NAACL-2000, Seattle Washington.

Pinker, S., 1994, *The Language Instinct—How the Mind Creates Language,* Harper Perennial.

Reeder, F. & Harper, L. 2000. Parlez Vous Francais? Multilingual Interaction in Virtual Worlds/Communities. In *Proceedings of Virtual Worlds and Simulation, VWSIM-2000,* San Diego, CA.

Schum, D. A., 1994, *Evidential Foundations of Probabilistic Reasoning.,* John Wiley & Sons, Inc.

Vauquois, B., 1968, A survey of formal grammars and algorithms for recognition and transformation in machine translation, *IFIP Congress-68* (Edinburgh), pp. 254-260.

Weaver, W., 1947, Letter to Norbert Wiener, March 4, 1947, (Rockefeller Foundation Archives).

Weiner, N., 1947, Letter to Warren Weaver, April 30, 1947, (Rockefeller Foundation Archives).

Winograd, T., and Flores, F. , 1986, *Understanding Computers and Cognition*, Addison-Wesley.

Biography

Ms. Reeder is a graduate student at George Mason University, completing her Ph.D. in Information Technology, focusing on evaluation of machine translation. By day, she is a Lead Artificial Intelligence Engineer for The MITRE Corporation. Her areas of work are in machine translation, multilingual data analysis, and enabling technologies. She is a developer of the CyberTrans system which puts MT (both commercial and research) in the hands of actual users.

[⊥] The views expressed in this paper are those of the author and do not reflect the policy of the MITRE Corporation.