Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2007 Proceedings

Americas Conference on Information Systems (AMCIS)

December 2007

# Semantic Machine Translation in Object Oriented Programming: Toward a Comprehensive, Idiomatic Model for Multilingual Communication

Philip Crowder
*Virginia Intermont College*

David Marlow
*University of South Carolina - Upstate*

# Semantic Machine Translation in Object Oriented Programming: Toward a comprehensive, idiomatic model for multilingual communication

**Philip D. Crowder**
**Virginia Intermont College**
philcrowder@vic.edu

**David W. Marlow**
**University of South Carolina – Upstate**
dmarlow@uscupstate.edu

## Abstract

*This research in progress reports on an investigation into a new approach to Machine Translation through semantic relationships as the primary pivot rather than a subsidiary element. This research will be of interest primarily to researchers in Machine Translation and those working with Object Oriented Programming as well as those with an interest in the semantic nature of lexical relationships.*

**Keywords:** *Machine Translation, Object Oriented Programming (OOPs), Semantics, Universal Networking Language*

## Introduction

Automated machine translation (MT) from one language to another has been a hotly pursued goal for more than 40 years (cf. Ledley and Wilson, 1962). Many people with translation needs now turn to websites like Babelfish.com which provides a word for word translation between many different languages. Babelfish and most of the many other free translation websites are based primarily on word lists which are cross-referenced between multiple languages. While this is certainly better than no translation capability, the result of this type of word for word translation is awkward at best and can easily result in a dramatic shift in meaning.

Human translation, in contrast, is an art of approximation and accommodation. Translators do not largely give word-by-word transliteration, rather they give the meaning, the gist, which is embodied in the individual groups or segments of words because transliteration, while successful at the micro level, fails to truly convey the full semantic load of the original discourse. In the computer language tests we have run most translations (between English, Korean, Chinese, Japanese and Spanish) have been similarly successful at the micro level but problematic in expressing true semantic equivalents.

To be specific, the popular Babel Fish web translation site has issues with very common greetings. In translating the standard Chinese greetings, the age/respect distinction between the Chinese 你好 (ni hao) and 您好(nin hao) is lost as both are translated to "hello". More distressingly, the Chinese questions "How are you?": 你好吗 (ni hao ma) and 您好吗 (nin hao ma) are rendered "you are good" without any recognition of the question particle 吗 (ma), thereby drastically distorting the nature of these conventional conversation starters from questions to statements. The English to Korean translation results are even worse. "How are you sir?" returns "너는 어떄요 의 각하?", which renders how as an adverbial construct, you in a highly informal or contemptuous form, and sir as a reference to a superior being. This type of confusion might be defensible with colloquialisms and idioms, but renders this type of tool dangerous when standard greetings are unreliable.

This paper outlines a work-in-progress that aims to determine if a semantic -based language translation system is feasible. The objective of the research is to develop a semantic-based translation methodology that would provide a hitherto unobtainable degree accuracy and flexibility in machine language translation through development of a dynamic Java-based class system. We have demonstrated above the inherent shortcomings of lexical-based translation systems and, in the following pages, introduce a semantically-based system. This first attempt with a small and theoretical sample shows sufficient promise to validate further exploration of the topic. If successful, a dynamic, accurate, user-friendly and ubiquitously available MT system could well rival the impact of Gutenberg's printing press. Applications for both business and personal expression are endless.

## Current MT Implementations

Modifications and improvements on word list correlations do, of course, abound. Word list matches have been enhanced by building in grammatical, statistical and semantic elements into choosing words most appropriate for the translation task at hand. In 1994, IBM introduced the Candide system which drew on statistical probability and complex algorithms to augment the application of translation by word list, with morphological and syntactic information (Berger, et. al). Seven years later, Yamada and Knight introduced a syntax-based translation system which, while building on the then conventional word by word statistical models, incorporated syntactic parsing trees to structure sentences in the target language. This enhanced comprehensibility of translated text by producing sentences with the target languages' default sentence structure and word order. It additionally provided a framework to manage particles like 吗 (ma) from the example above. Perhaps most importantly, they also sought to incorporate phrase-based translation as a complement to the word by word method (2001).

The phrase translation model has received significant attention (cf. March & Wong, 2002; Koehn, Och, and Marcu, 2003; Venugopal, Vogel, & Waibel, 2003; Zens & Ney, 2005; and Pang, et al, 2005). Counter-intuitively, this phrase-based translation employs phrases "composed of a series of words that perhaps possess no syntax or semantic meanings" (Pang et al, 2005:1). The composition of phrases is determined by statistical probability without regard for natural collocations or semantic units in either the originating or target languages.

Although semantics has been a concern of information specialists for more than 40 years (cf. Ledley and Wilson, 1962), progress toward incorporating semantic values and variables into machine aided language translation has been slow. In the early 1980s, Hirst (1983) put forward a foundational conception on which semantically based MT could be built. Seven years later Carasik and his colleagues again called for a computer language system with "a semantic orientation rather than a strictly syntactic orientation" (1990). In 1996 the Universal Networking Language (UNL) project was born at the United Nations University in Tokyo. The stated mission of this project is "to provide the methods and tools for overcoming the language barrier on the World Wide Web in a systematic way" (Introduction). The vehicle by which this is to be accomplished is a meta-language based on semantic principles and built on a layered model consisting of binary relationships, 'universal words' and attributes of those words.

The concept of a 'universal word', can be illustrated using the Meaning Triangle (Carasik, et al, 1990:30). The Meaning Triangle (Figure 1) illustrates the relationships between concepts, referents and symbols. The goal of all communication is to use language (SYMBOL) to express a CONCEPT which relates to the manifestation of the concept in the physical world (REFERENT). This is no small task in a monolingual environment; one person's prototypical conception (REFERENT) of "tree", for example may be an oak, while another may relate the same word to a mental picture of an evergreen. In the UNL a 'universal word' is a concept so basic that all human languages can be assumed to share the same underlying CONCEPT – REFERENT relationship. "Person" and "thing" are two examples. These 'universal words' are then further defined by a set of attributes. For example, a PERSON might have the attributes of male or female, young or old, foolish or wise etc.

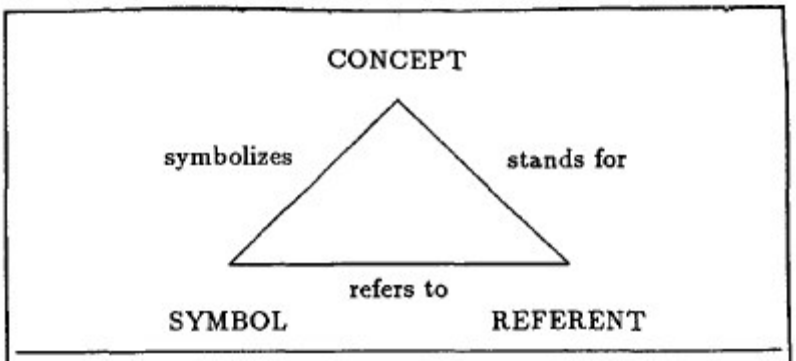



Figure 1: The Meaning Triangle
Different cultures have different written and spoken symbols for the same experiences.

## Semantic MT in OOPs

This paper lays the foundation for furthering the concepts of statistical word lists, phrase-based translation and the attributive definition of 'universal words' through the development of hierarchical classes of meanings modeled on the object-oriented programming structure. Similar to the UNL, our model would derive language-specific words as values or implementations of those meanings, but would do so only on the lower level of the hierarchical structure. A principle differentiator between our construction and that of the UNL is that in our Semantic MT in OOPs (sOOP) model, semantic properties are proposed at the phrase level as well as for individual words rather than for individual words alone. We submit that semantic classes correlate naturally with the hierarchical structure and implementations of Object Oriented Programming languages (OOPs). These semantic classes can be used to meaningfully represent linguistic properties of both words and phrases. While this is not a new concept, Ledley and Wilson referred to the efficiency of "object-language code" for MT as early as 1962 (p. 154), we have uncovered no research discussing a model which exploits OOPs concepts at both the word and (syntactically and semantically meaningful) phrase levels.

Object-oriented programming such as C or C++, Java and Visual Basic is a way of modeling objects, their associated properties (or attributes) and actions as they are acting or acted upon as a unit. A nominative, for example, may be instantiated as an object. A property may conveniently be thought of as analogous to an adjective that further defines the nominative. Actions are verbs, active or passive. A class is a generic grouping or representation from which an object is derived. Classes may be subsets of other classes in an "is a" relationship and these classes obtain all the attributes and actions of the parent class through a process called inheritance, and these methods and characteristics of the parent class may be modified by overriding them in the child class.

Let us take a hypothetical case that is typically used to illustrate OOPs and then take the same concepts and apply them linguistically. There is a group of objects instantiated as John Doe, Jane Doe and Sally Smith, which may collectively be described as the class, STUDENT. STUDENT has certain properties, or characteristics, associated with it. One such association is name, another address, and others phone number, ID number, class level, major, grade point average. Each of these properties has values associated with it. One would note that these characteristics would exist in common with any society which had a collection of entities which could be typed as "STUDENT". From this object class STUDENT there are the specific instances previously alluded to including the specific instances of students John Doe and Jane Doe. Each instance of STUDENT class has the same properties and actions associated with it. Each instance of STUDENT, be it Jane or John, however, has specific "values" associated with each property associated with each instance of the class student . An instance of John Doe of class Student would have property value pairings of address = "123 Munchkin St, Oz"; phone number= "1-800-Munchie"; and major = "culinary arts and cheerleading", for example. John would thus be implemented as a Cheerleading culinary arts major who happens to live in Oz and is, inferentially, a Munchkin. Jane Doe, whose properties come from the identical class student that derived and implemented John, nonetheless has different values associated with these properties such that a Jane Doe when instantiated, is a Senior majoring in Microbiology and has a GPA of 2.6. Though we could have also included her phone number and address such properties values were not salient for this Jane implementation. Note that we could not have included her height in this implementation of Jane (or John) for height is not a property defined or associated with the student class. You may only utilize the characteristics common to each class though it is not required that you implement all properties of each instance of an object derived from student.

Semantic groups are representations of objects that have certain properties and actions associated with them. These groups, being vessels or containers of intent or meaning, are represented in our model as class types (semantic concept groups) that have meaning that is common to every spoken language but which is implemented differently in each language. In other words, not all the available properties of each word meaning need be manifest at each implementation. The existence of properties provides the necessary selectivity and versatility for semantic concept groups and ensures cross-linguistic syntactic integrity.

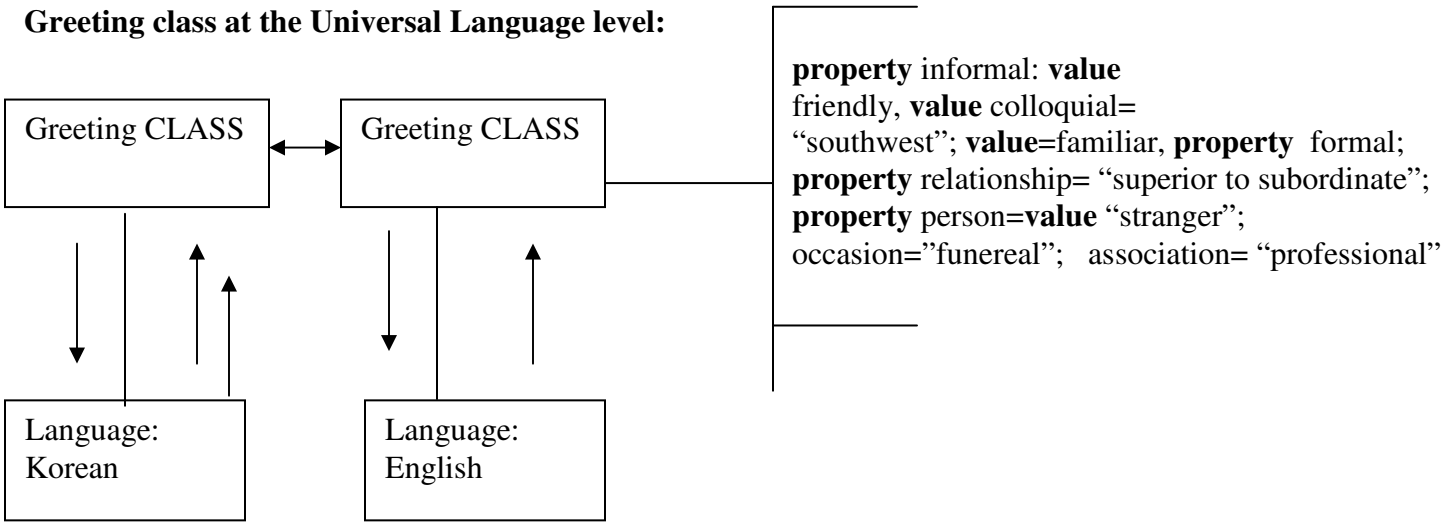
Thus, in a similar fashion to student we could have the linguistic construct of Greeting as a class. Implementations of the semantic concept annyunghaseiyo, for example, within the GREETING class would have certain properties universally related to greetings such as formal, informal, friendly, familiar and so on. Then, through manipulation of the values associated with the properties of the root concept of여보세요 (annyunghaseiyo) would be derived such implementations as hi, hello, how are you depending upon the values associated with those properties. Only after the various properties and associated values had been selected would specific terms from specific languages be derived.

Table of class GREETING with sample properties, values and derivations.

| Semantic concept | Property | Value | Derivation |
|---|---|---|---|
| Greeting | Formal | Subordinate to super | Hello |
| | Informal | Associate | Hi |
| | Informal | Colloquial | Hi, how are you |
| | Formal | Professional | Good to see you again |
| | Informal | To child | Hey [name] |

While the table gives a static view of relationships what follows is a diagrammatic representation of the flow of data from source language to destination language where the word concept Hello is first encapsulated as greeting class, with the appropriate associated properties and values, then sent to the destination language where those word properties and values are implemented appropriately in the destination language. This avoids the inaccuracies involved in word on word lexical interpretation.

**Greeting class at the Universal Language level:**




Indeed while such expressions as hi or how are you may not be linguistically mappable by word-list associations the correct language specific values may be easily derived from the class model described above. Meanings may be accurately represented as attributes of an object of the greeting class, with properties of informal: friendly, colloquial= "southwest"; familiar, warm or formal: relationship= "superior to subordinate"; person= "stranger"; occasion="funereal"; association= "professional" etc.

Only at the class-concept language-implementation level , where instances may be derived from general concepts, would word associations be made to a specific language from defined properties and values in a manner analogous general programming code being interpreted into a specific machine language. For example, class Greeting with the attribute value of "subordinate to superior" would be rendered as 안녕하십니까? (annyunghapshimnikka?) While an attribute value of "child" would derive 안녕 (annyoung).

Furthermore, though the hierarchical structure of our sOOP model, is immutable, the classes of objects themselves have the versatility of mutability, not only with the addition or deletion of property strings and alteration of value lists, but classes may be derived from other classes and classes themselves may be deleted. This allows the necessary flexibility at the language level to bring about the correct meaning with the correct word usage at the specific language level irrespective of the sending and the receiving language.

## Next Steps

In this paper we describe the first step and general framework for a semantically based MT translation system. We recognize that much work remains to be done. Our next step will be to formalize a working model codified in JAVA script using the greetings class. We will populate the same for Chinese, Korean, Hindi, Spanish, and German using native speakers of these languages as informants. We will then test the system with other native speakers of these languages via a matrix which compares English-German (which are structurally related) and English-Spanish (with a high density of shared cognates) with the less related pairings of English-Chinese, English-Hindi, and English-Korean. The system will be refined based on feedback from informants and retested. When English-to-other language translations are fine-tuned, other pairings will be considered (e.g. Chinese-German, Hindi-Korean) and the system will be fine-tuned again. When this step has been completed, we will expand our class base to simple notional concept classes such as "shopping" and "working with computers" and retest the model, again refining based on our findings.

As a part of the steps above, we will explore specific protocols for populating attributes of classes. At this point we are considering a model based on the systems currently used to develop word and phrase lists, and tag corpora for parts of speech (cf Biber 1988 and Biber, et al, 1994).

## Summary and Conclusions

We have presented here the first stages of a model that will extend the core concepts on which MT is founded. We propose a model of MT translation that will include commonly collocated phrases which are meaningful at both the syntactic and semantic level and which are tagged with attributes modeled after OOPs programming. This will allow for a level of nuance to be translated which is unattainable with the prominent models currently described in the MT literature.

As with any first stage model, there is much work remaining. Some of the questions which are to be addressed more fully include the reliability of a system for populating the class attributes across languages, the level to which phrases can and should be used in preference to word level association, and the cost in terms of processing power and speed to accomplish sOOP translation. It is hoped that this first exploratory examination of this new concept will provoke additional questions and eventually result in a ubiquitous system for translation that will be accurate on both the micro and macro levels.

## References

Biber, D. 1988. Variation Across Speech and Writing. Cambridge University Press.

Biber, D. Conrad, S. and Reppen, R. 1994. Corpus-based approaches to issues in applied linguistics. Applied Linguistics. 15(2): 167-189.

Introduction of the UNL. Online document: http://www.unl.ru/introduction.html. Accessed May 2, 2007.

Koehn, P., Och, F. J., and Marcu , D. (2003) Statistical Phrase-Based Translation. In Proceedings of the Joint Conference on Human Language Technologies and the Annual Meeting of the North American Chapter of the Association of Computational Linguistics.

Ledley, R. S. and Wilson, J.B. (1962). Automatic-programming-language translation through syntactical analysis, Communications of the ACM: 5:3:145-155.

March, D. and Wong W. (2002). A Phrased-Based, Joint Probability Model for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP.

Uchida, H. and Meiying Zhu. 2001. The Universal Networking Language
beyond machine translation. In International Symposium on Language in Cyberspace.

Venugopal, A., Vogel, S. and Waibel, A. Effective Phrase Translation Extraction from Alignment Models. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. 319-326.