Building a Controlled Natural Language Framework for Real-time Machine Translation

Sebastián León Calderón
Intel, Costa Rica

Abstract
This study presents a framework that addresses the need for real-time communication between persons without a common language through the use of Controlled Natural Languages (CNLs). The framework consists of a set of CNLs (initially based on English and Spanish) where text compliant with any one CNL is mapped to an internal, unambiguous, language-independent representation which, in turn, is automatically translated into valid text in a target natural language that not only retains the meaning of the original text but also appears natural to native speakers. In addition, the study proposes a software system similar to modern instant messaging clients to effectively facilitate said communication.

Key words: controlled natural language, machine translation, Grammatical Framework, instant messaging, natural language processing

Resumen
En este estudio se presenta un marco de trabajo dirigido a la necesidad de la comunicación en tiempo real entre personas sin un lenguaje común, mediante el uso de los *Controlled Natural Languages* (CNLs). El marco consiste en un conjunto de CNLs (inicialmente basados en el inglés y el español), en el cual el texto sometido a cualquiera de estos CNL es proyectado hacia una representación inequívoca de lenguaje independiente que, a su vez, es automáticamente traducida a un texto válido en un lenguaje meta natural que no solo conserva el significado del texto original sino que aparece como natural a los propios hablantes nativos. Además, en el estudio se propone un sistema de *software* similar al sistema moderno de mensajería a los clientes para facilitar dicha comunicación.

Palabras claves: lenguaje natural controlado, traducción por máquina, marco gramatical, mensajería instantánea, proceso de lenguaje natural

Recepción: 19-9-13 Aceptación: 23-4-14

Introduction

Machine translation (MT), defined as the use of computers to provide automatic translation from one language to another without human intervention or assistance, is made far from trivial by the ambiguity inherent to natural languages. Such ambiguity, while allowing for expressiveness and interpretation (both essential in human communication), comes at the expense of exactness, a most desirable feature in the context of automatic and lossless processing of text: the first (and most complex) of the two primary steps that constitute the machine translation task (the other one being the mapping of the output of such processing to text in the target language).

A well-known approach towards reducing said ambiguity, and the chosen method for this paper, is the use of controlled natural languages (CNLs) which can be broadly described as well-defined subsets of specific natural languages often aiming to either aid communication for non-native speakers or allow for automatic extraction, processing and representation of knowledge presented as text. CNLs tailored around the former aim (commonly referred to as human-oriented CNLs) are often "true" subsets of their source natural languages in the sense that they merely define a number of restrictions on which grammatical structures and lexicon are allowed. On the other hand, those made with the latter goal in mind (and fittingly referred to as machine-oriented) tend to be defined as formal (and thereby unambiguous) languages whose grammars mimic the ones of their source languages, while not strictly being subsets of them, but still retaining as intuitive a readability as possible.

As computer technology progresses and becomes more of an everyday commodity (especially through always-on, always-connected devices such as smartphones and tablets), and information becomes much more readily accessible, the need for communication between people also becomes that more urgent, and is expected to be made progressively easier. Instant messaging (IM), in fact, has become one of the most popular means of communication for its cheap cost, immediate nature and relatively low time investment. However, most modern instant messaging clients do not provide much support as far as enabling communication across languages goes; instead, they rely on the users' ability to agree on a common language. While this might be appropriate in most everyday situations, it can be a prohibitive factor when such a common language does not exist. A partial solution to this problem would be to employ an automatic translation tool (such as Google Translate) for each instance of communication, a task that not only is bothersome but also known to be failure-prone, as such tools do not guarantee accuracy in translation and do not inform users as to what will and what will not be precisely translated. This situation becomes even more problematic when more than two users are considered, as communication reliant upon imprecise tools can quickly lead to confusion and accidental disagreement.

The research presented herein aims to attack the problem of cross-language instant communication by replacing the original languages of users with a machine-oriented, formal CNL based upon it (e.g. Controlled Spanish, Controlled English) and thereby significantly simplifying the implicit task of machine translation. This work intends to provide a computer system that facilitates said communication for as numerous a set of natural languages–and thereby as broad an audience–as possible and is structured as follows: Section 2 gives an overview of the task of machine translation, is mostly concerned with the many difficulties it faces; Section 3 makes an argument about the usefulness of CNLs when applied to said task, and presents existing technology for the implementation of CNLs; Section 4 presents a framework that utilizes said technologies to facilitate sentence-level machine translation; Section 5 proposes a specific application of this framework in the form of an IM system; Section 6 provides a summary of the study and an presents an outline of future work and interesting applications.

The Problem of Machine Translation

As touched upon before, translating a text is a very complicated task to automate. Particularly, the subtask of deriving meaning from text (which to human translators in possession of the required familiarity with the language is the most trivial) can quickly become a gargantuan enterprise, even when the source and target languages are not terribly distant from each other. In the following paragraphs, several of the most immediate complications one faces while designing a machine translation program are listed as a means to prove the previous claim, and also to provide the reader with perspective for the evaluation of the CNL-based approach to be presented later on.

Let there be, for the purposes of the current argument, a trivial MT system that simply parses the source text one word at a time, looks up a fitting translation in a dictionary for each one, concatenates them, and gives the result as an output. Three problems are immediately evident: some words may not have an equivalent in the target language; some words may have more than one possible meaning (and thereby multiple candidates for translation); and the grammatically correct order of the words may differ from one language to the other. Of these, the former is the only one that has a plausible (albeit far from ideal) solution at word-level: replacing the word with a set of words that approximate its meaning as closely as possible. As for the latter, no decisions can be made unless the context is taken into consideration, as the meaning of a specific word is often largely determined by the sentence it belongs to (or even adjacent sentences), and the ordering of words evidently cannot be determined unless the full sentence is analyzed at once.

As a consequence, it becomes necessary to modify the MT system so that sentence-level analysis is performed. Such upgrades would allow the system to take one sentence and identify the morphological functions of its words in an attempt to produce correct ordering of words (and disambiguate some words). New problems arise, including but by no means limited to: some specific word combinations may have more than one possible meaning (idioms being particularly noteworthy); some loosely-defined combinations of words may introduce meaning that differs from what's otherwise expected (as is the case of collocations); the specific meaning of some words may not necessarily be determined by their morphological function within the sentence; and–perhaps most aggravatingly–some sentences might possess more than one applicable mapping of words to morphological functions.

At this point it becomes apparent that, in order to achieve accurate translation by means of analysis of the text only (as opposed to the use of statistical methods), it is necessary to follow one of two approaches: deriving meaning from the text (which requires some sort of ontological representation), or designing a mechanism for the identification of well-known occurrences within the source language as well as their mapping to grammatically correct phrases in the target language. The former approach, while closely resembling human translation, is a very complex task to achieve and requires a knowledge base to describe all possible ontologies and is thereby limited to the exhaustiveness of said database. The latter approach, while avoiding the need for ontologies, depends on the creation of a large and complex set of rules that address every possible occurrence: a daunting task to say the least.

It must also be noted that natural languages are not static; they evolve over time as they are constantly enriched with calques, foreign and loan words, and lexical and syntactical developments of their own (many of which take place at a local scope). An argument can thereby be made that a MT system can never be "complete" in the sense that it cannot be expected to provide a plausible translation for every possible text at any given point in time. The most pressing concern, however, is the relative complexity of introducing new grammatical structures into the system, regardless of the chosen approach: an ontology-based system would need to be cross-checked so that new concepts do not conflict with existing ones, whereas a rule-based one is likely to require serious revision to ensure that no confusion comes from potentially overlapping structures. In other words, as long as the source language is taken as a whole, the complexity of designing a truly comprehensive and accurate MT system will be almost prohibitively elevated.

Controlling the Language

Following the previous argument that an MT system would benefit from limiting the grammar and lexicon of the source text rather than allowing the full language to be utilized. While this technique would irremediably cause the system to be "incomplete" as defined before (which, the author has argued, is already implausible), it would allow for it to be "complete" within the limited scope. It must also be noted that the flexibility of natural languages more often than not allows for notions to be represented in several different ways, which should allow for a reasonable amount of source text to be adjusted so that it adherers to the restricted grammar and lexicon.

As touched upon before, a CNL is precisely, the result of restricting a natural language in order to reduce its ambiguity; a goal that can be achieved in a top-down (successively banning undesirable features) or bottom-up (start from scratch and successively adding to the CNL) fashion. While a top-down approach might be suitable for human-oriented CNLs mainly concerned with keeping communication simple but not necessarily unambiguous, it does not align well with the goals of machine-oriented ones, where formality is essential, as it would mean reverse-engineering the CNL from its source language, a task that requires a very intensive up-front design stage. Instead, the bottom-up approach of iteratively adding grammatical structures and lexicon allows for cyclic implementation, where each loop contains a smaller design stage, and the implementation and testing of said design. In the context of MT, the more restricted and close to formality the source text, the easier it will be to provide a high-quality translation for it. Thus, the bottom-up, machine-oriented approach to creation of CNLs is a natural choice.

Indeed, much previous work exists in this area, with several machine-oriented CNLs having already been implemented and even applied to the task of MT. An excellent, well-known example is Attempto Controlled English (ACE): a CNL with a formal grammar based on first-order logic and originally designed for software specifications and later expanded upon for use in automated theorem proving with human-readable inputs and outputs, as well as interoperability with the languages of the Semantic Web. What makes ACE remarkable in the area of MT, however, is its application as part of the AceWiki-GF project: an effort to create a multilingual Wiki environment. In AceWiki-GF, users can add knowledge to the wiki as sentences in one of several supported (albeit heavily restricted) languages, and these sentences are automatically mapped to ACE and later translated from ACE into the remaining supported languages so that the added knowledge is simultaneously made available in all language-specific versions of the Wiki at once.

The “GF” in AceWiki-GF refers to Grammatical Framework, a programming language that permits the implementation of custom grammars for formal languages, and is specifically optimized to support features inherent to natural languages, going so far as to provide a Resource Grammar Library (RGL): a predefined library that describes a broad set of grammatical structures for about thirty different languages, collaboratively built by its users. The AceWiki-GF project relies on this library to provide translation into and parsing from languages other than English, and employs a GF-based implementation of ACE to provide a language-neutral representation of all knowledge stored in the Wiki.

AceWiki-GF, as a knowledge-oriented project, imposes a number of additional restrictions to the language that, while allowing for automatic reasoning and interoperability with the Sematic Web, make it unsuitable for use as a reliable MT system whenever source text does not describe knowledge. Specifically, all text within AceWiki-GF takes on the form of either a fact, such as “Spain is a country”, or a question over previous facts, such as “What is a country?” (the answer to which, as automatically generated by the Wiki, would be a list of all countries previously entered through statements analogous to the previous example). Human communication, however, lacks this sort of structure and it is therefore unrealistic to expect source text to adhere to said restrictions. A more flexible approach is thereby required in order to provide a MT system that's useful for human interaction.

A Controlled Natural Language Framework

This work addresses the need for multilingual human communication through the proposal of a CNL-based multilingual MT framework. This is, a reusable software library that can be utilized in the implementation of concrete MT systems. One particular proposed implementation is the instant messaging system with built-in real-time translation discussed in Section 5. Said framework is composed of a set of source languages and a set of target languages, which may overlap but do not necessarily match. The input for the framework is some text written in any one of the source languages, while its output is the corresponding translation for each target language.

In order to begin the design stages of the framework, it is first necessary to determine exactly what is to be controlled. From the discussion in the previous section it is evident that every source language needs to be controlled; the target languages, however, might not. The decision depends on whether or not it is desirable to have biunivocal translation; this is, that for any given text in a source language, the translation produced by the system can be fed back to it in order to obtain the original text. Biunivocal translation would require that the target language be controlled as well (effectively causing the source and target language sets to be identical), while a "translate-and-forget" system would instead benefit from having the full target languages available to it, as this would allow for greater specificity in each individual translation. Given that the goal in mind is facilitating human communication, it is preferable to provide natural-sounding translations even though they may not in turn be used as inputs. Thus, source languages are controlled, but target languages are not.

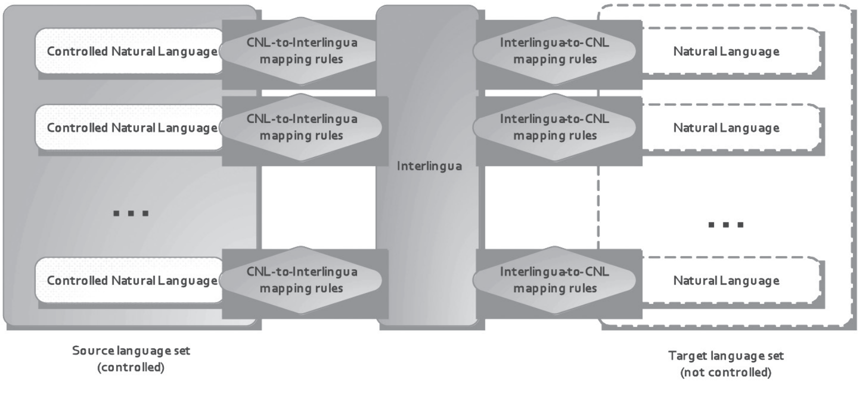
As touched upon in Section 2, the most intuitive approach towards providing automated translation from a source to a target language would be to design a set of rules that define how specific syntactical structures and morphological paradigms are to be mapped. Applying this direct mapping to a multilingual environment would necessitate that, for each pair of languages, a set of rules is designed and implemented. While this is definitely feasible for a very small number of languages, it scales very poorly as languages are introduced into the framework, since each new source language would require a set of rules to be added for each target language and vice versa.

Traditional human translation techniques, on the other hand, typically involve an expert that is familiar with both source and target languages and is capable of deriving meaning from the text and producing a translation that approximate that meaning as closely as possible. In a multilingual environment, and provided that the expert is familiar with all applicable languages, this remains largely unchanged, as the meaning and not the exact wording is what survives the translation process.

In the context of MT, any meaning extracted from text needs to be stored as a machine-friendly representation, namely an ontology. Ontology-based translation employs a set of rules that describe how the ontologies are produced from the text, and a second set of rules which define how the ontologies are transformed back into text (this is, the translation). It should be noted that this not only closely resembles the human translation process but also does not cause the framework to suffer from the poor scalability of the direct mapping approach, as incorporating a new language to either set calls for the addition of only one extra set of rules.

From the previous reasoning the author concludes that the ontological approach should be preferred for the framework. Specifically, ontologies take the form of a language-independent representation of text (or rather, meaning that can be expressed as text) tentatively called Interlingua. As presented in Figure 1, the complete framework is formed by the aforementioned source and target language sets, the Interlingua, and the rules that allow for mapping between languages and the Interlingua. It must be noted, however, that the target languages are pre-existent and do not need to be designed; this is not the case with the source languages which, as CNLs, require to have a grammar defined for them. As a consequence, target languages are not actually built into the framework, as evidenced by the dotted lines in Figure 1.

Figure 1
Conceptual look at the framework


Controlled Natural Language
CNL-to-Interlingua mapping rules
Interlingua-to-CNL mapping rules
Natural Language
Controlled Natural Language
CNL-to-Interlingua mapping rules
Interlingua-to-CNL mapping rules
Natural Language
Interlingua
...
...
Controlled Natural Language
CNL-to-Interlingua mapping rules
Interlingua-to-CNL mapping rules
Natural Language
Source language set (controlled)
Target language set (not controlled)

The Interlingua (middle) plays a central part in the translation process, being mapped to from the source controlled languages (left) and mapped into the target ones (right). Target languages are technically not a part of the framework, but the rules for mapping into them are.

The addition of the Interlingua to the framework effectively causes source and target languages to be fully abstracted from each other and splits the task of translation into two much simpler subtasks: *parsing* the original text to a representation in the Interlingua, and *linearizing* said representation into the translated text. As a consequence, it is the Interlingua which ultimately defines what controls need to be introduced to each source language in order to allow for it to be plugged into the framework, as well as the nature of all mapping rules.

Inspired by the AceWiki-GF project (which in turn uses ACE as an Interlingua of sorts), the implementation of this framework will extensively rely on Grammatical Framework for the definition of the grammars required by the Interlingua and any CNLs to be used as source languages. Given than GF already provides mechanisms for both the parsing and linearization stages of the translation process, design will focus on successively adding syntactical structures to the Interlingua and accordingly building upon the CNLs so they make use of said structures. As the author's linguistic expertise is mostly limited to English and Spanish, these two will be used as starting languages on both the source and target ends, with others to be added at later stages of development.

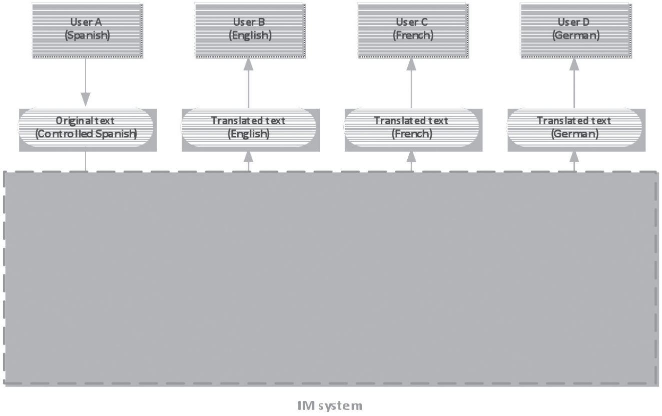
Using the Framework for Real-time Translation

While the framework presented in the previous section has a great number of potential applications, the main goal of this work, which eventually lead to its conceptualization, was the provide a means for real-time communication between two or more people in scenarios where no common language is shared by all the participants and no human interpreter is at hand. While providing a system that could take dictation from users and produce a translated voice feed would certainly be ideal, proper usage of such a tool would depend on the user very meticulously composing his sentences so that they adhere to the restricted grammar of CNLs; for the average user, this is more likely to result in an unpleasant experience rather than a convenience. In addition, the technology required for extraction of actual speech from audio recordings belongs to a much broader area of research and thus exceeds the scope of this work.

Instead, this work proposes an instant messaging system. A typical usage scenario, presented in Figure 2, would involve an user typing text that complies with the grammar of a CNL based on their chosen input language and then prompting the system to deliver the message to one or several other users; these users would each receive a translated version of the message corresponding to their particular language selections. Neither the sender nor the receivers need to have prior knowledge of any language other than their own, as the system handles the entire translation process seamlessly. Instead of attempting to find a common language for communication, the only requirement is for users to previously learn what restrictions the CNL imposes and adjust their word selections accordingly during usage.

The writer notes how unlikely it is for even the most experienced users of such a system to consistently write well-formed sentences. In order to alleviate the frustration that may come from having inputs rejected by the parser, the system should also include the ability to provide real-time feedback to users regarding their selection of words, in either predictive or reactive forms. Predictive feedback constantly parses the input and provides the user with a list of words that can follow the most recently typed one,

Figure 2
Typical usage scenario for the IM system



User A, a Spanish speaker, submits some text in Controlled Spanish. The system seamlessly translates the text into English, French and German for users B, C, and D, respectively. Internally, the original text is parsed into the Interlingua and linearized into the target languages.

where applicable. Reactive feedback attempts to parse the text on submission and, upon encountering a parsing error (this is, a word that makes the sentence invalid), it prompts the user with an appropriate error message. Each user should then choose whichever type of feedback allows for the more enjoyable experience, according to their personal perception. Another very desirable characteristic for the tool would be to have the produced translations shown not only to their intended addressees, but also to the composer of the original text. The addition of such a feature, not only is communication achieved, but also some learning of foreign languages could be done to be used in future instances of unassisted communication.

Discussion and Future Work

The present study has yielded an argument that the task of machine translation without the use of statistical methods suffers from the elevated complexity and ambiguity present in natural languages. It presented CNL technology as a means to alleviate such difficulties by restricting the grammar to a “controlled” subset, and provided an example of said technology being successfully applied to MT in the form of the multilingual semantic Wiki known as AceWiki-GF. The study noted, however, that AceWiki-GF’s knowledge-oriented paradigm makes it unsuitable for use in the translation of regular human communication. As a result, the writer presented a concept for a MT framework based on CNLs and a language-independent ontology tentatively dubbed Interlingua. Said framework permits translation of CNL-adherent text to many natural languages, requiring only that mapping from the CNL to the Interlingua, and from the Interlingua to the target natural languages be provided. Finally, the author proposed the design of an instant messaging application that utilizes the framework to provide real-time communication between users in a multilingual setting, and gave rationale for the need of a feedback module at the input validation stage.

The writer highlights that the presented framework is currently moving from the conceptual stage to the initial iterations of design, and much work yet needs to be done before a prototype can be presented. However, he also notes that much of the underlying technology is already provided in tools such as Grammatical Framework, allowing for a rather speedy implementation once the starting design has been completed. In addition, a number of open-source IM systems have been made freely available in the Internet, which should allow for both the framework and the IM client to be developed simultaneously.

Future work in this area, as touched upon in Section 5, may be directed towards the addition of a voice interface to the proposed IM system. While this may not be practical for new users who can be expected to regularly fail to produce CNL-adherent sentences, it could potentially speed up the process of submitting text for translation for more advanced users. On the receiving end, proving an audio feed rather than (or as well as) translated text is just as advantageous, regardless of the user's familiarity with the system. Finally, and given the implementation-independent nature of the framework, it could potentially be used as the main translation mechanism for a system in any other multilingual setting, such as micro-blogging, news feeds, online technical documentation, etcetera.

Bibliography

Kuhn, T. (2013). A Principled Approach to Grammars for Controlled Natural Languages and Predictive Editors. *Journal of Logic, Language and Information*, 22(1), 33-70.

Kuhn, T., & Kalijurand, K. (2013). A Multilingual Semantic Wiki Based on Attempto Controlled English and Grammatical Framework. *Proceedings of the 10th Extended Semantic Web Conference*. Berlin: Springer.

Ranta, A., & Angelov, K. (2010). Implementing Controlled Languages in GF. In Fuchs, N. E. (Ed.), *Controlled Natural Language* (82-101). Berlin: Springer.

Schwitter, R. (2010). Controlled Natural Languages for Knowledge Representation. *Proceedings of the 23rd International Conference on Computational Linguistics*: Posters. Stroudsburg, PA, USA: Association for Computational Linguistics.