Multilingual Authoring through an Artificial Language

Marcos Franco Sabarís, Jose Luis Rojas Alonso, C. Dafonte, B. Arcay

University of A Coruña
C/ Zaragoza 13, 5C
36203 Vigo
Spain
mfs@mundo-r.com

Abstract

Nowadays, there is a growing need for dissemination of documents in several languages. Machine translation is usually regarded as a possible solution for this, but so far it cannot provide acceptable translations of unedited texts. Several methods which involve human participation in computerized processes of translation have been proposed, but none has given really satisfactory results (except in some restricted contexts). In the UTL (Universal Translation Language) project, which we present here, we propose a new approach to multilingualization, based on the usage of an artificial unambiguous human language in which the human translator writes the source text, and then gives it to the machine to translate into other languages. The nature of this constructed language, which is optimized for this role, ensures the high quality of the results rendered by the computer.

Keywords

Multilingual, authoring, human-assisted, artificial language

Introduction

Since the days when the computer was invented (and even before), it has been a dream shared by many people that one day machines could perform the task of human translators, that is, to convert one text from one language into another. However, machines lack (and will probably lack for a long time) the intelligence necessary to carry out efficiently such an assignment. Nevertheless, it seems to be perfectly in accordance with the state of current technology in this field to expect the machine to translate one text to another (or several other) language(s) with the assistance of a human operator. This possibility has been tried for example under the form of dialogue-based systems for monolingual authors in the past (Boitet and Blanchon 1993), whether by helping the computer in the conversion to the target language (in direct systems), or in the conversion to the intermediate language (in interlingual systems). Pre-edition, with the introduction of a controlled language, has been another method employed to make high quality machine translation possible under some circumstances (Pym 1990). In this paper, we want to present a new approach to multilingual authoring to which we have given birth in the confluence of two seemingly unrelated disciplines: computer-aided translation and interlinguistics or constructed auxiliary languages. This new approach takes form in the UTL (Universal Translation Language) project.

UTL: a new method of multilingualization

Born as a consequence of a growing need in today's world, the aim of the UTL project is to provide a tool to convert a specific text written in a given natural language into an indefinite number of other languages, with the help of a computer, in a process of human assisted machine translation. The role of the human translator involved in this process will be confined to provide the computer with a translation of the original text into a special artificial language (the UTL language) that the computer can "understand" and translate better than the original text (written in a natural language). The UTL language is therefore a constructed human language, based on Esperanto, which has been optimized for being processed accurately by a translating software, and which is to be employed by a UTL human translator who has previously been instructed in it.

Basically, the translation process proposed in the UTL system consists of three steps:

1. A human translator converts the source text (written in any language, Russian, English, Japanese...) to the UTL language. Needless to say, the text can also be directly authored in UTL.

2. The text in UTL is given to the translation software, which will be prepared to analize it and to generate versions in one or more natural languages.

3. Generated translations can be used as they come, or they can be post-edited if there is a need to enhance the style.

In this way, the system allows us to obtain, at the cost of just one manually made translation (from natural language to UTL), versions in several languages with a final quality noticeably superior to conventional automated translations (i.e. from natural language to natural languages). Thus, an expert in UTL is capable of translating a given text into fifteen, twenty or more languages, in the same period of time one would normally spend on one single manually made translation.

In the different sections of this paper, we will give a general description on how the UTL system works and how it can be employed in an interlingual MT system. We will also explain its relation to currently existing technologies and to other artificial languages that have been created in the past.

The Present Situation of Translation Technologies

In the early days of machine translation (about the middle of the 20th century), investigators had high expectations that fully automated high quality translation (FAHQT) between natural languages would be attained in a few years. Today, fifty years later, with a computing technology which could not even be imagined by the pioneers, few experts believe in this possibility. The problem seems to be more in the nature of human languages than in the power of computers. Even if not obvious for the average speakers, the languages we use to communicate with each other are full of ambiguities, imprecisions, idiomatic utterances, anaphora, ellipsis, etc, elements which do not bother human intelligence, with its knowledge of the real world, but which give machines a hard time at the moment of analizing a text in one language and converting it into another.

As a result of this, the scope of machine translation points nowadays to less ambitious goals, though not unimportant. Some of them consist in rendering draft-quality translations which can be useful to give the reader a gist of what the original text says (an application which has found considerable popularity among Internet users through on-line translators), or providing the human translators a first version which they can then post-edit, an option which can save them time and effort with certain types of documents (technical, repetitive...). An interesting possibility, implemented in some industrial systems (e.g. Systran in Xerox), is that of using a controlled language (a simplified version of a natural language) to write the source document, which makes the output of the translation system more reliable. This technique is useful and cost-effective with texts of technical contents when they are meant to be translated into several languages. The results become better still when the vocabulary and the grammar used in the source language can be restricted more dramatically, as in the case of weather reports (cf. Météo system), and so we are talking about sublanguages.

The philosophy of the UTL project is similar to that of controlled languages in that we make the machine translate from a language which due to its grammar and lexicon is easier to analize automatedly. This strategy is carried in UTL much further than current controlled languages, as we propose the usage of a new constructed language particularly suitable for this purpose, instead of getting by with an impoverished version of a natural one. This will not only make even better translations (near to error free and enabling more complex syntax as well as larger basal lexicons) possible, but will also widen the scope of this procedure to general domain documents.

UTL as an Artificial Human Language

Since the 17th century, there have been many projects of artificial languages, most of them aimed at international communication or at philosophical purposes. Nowadays, with the importance of machine translation, an artificial language which is easy to learn and which expresses ideas in a way similar to most major natural languages but at the same time in a precise and unambiguous manner comes in handy, in the way proposed by the UTL project. UTL's design is inspired by the most developed auxiliary languages, with special regard to Esperanto as a living example of an artificial language that works efficiently in practice. We have chosen Esperanto among the several existent artificial languages as it is the most developed of them all. It has complete dictionaries and grammars, and has been used as a second language by a community of hundreds of thousands of speakers around the world for more than a century (Janton 1976: 11-32). Many of the characteristics necessary for a *translational* language like UTL are already present in Esperanto, though a few new features have been incorporated into the language in order to optimize its unambiguity and semantical capabilities. "Translational language" is the utterance we have chosen to make reference to this new role a language can play in the field of computational linguistics.

On the other hand, if we have defined UTL so far as a special kind of controlled language, our *translational* language could also be defined as an MT intermediate language (interlingua) aimed at a direct authoring in it (and not just for internal use of the computer). This entails a formal difference from traditional interlinguas (with their characteristic symbolic list representations) in that a translational language must be preferably human rather than electronic, (relatively) easy to master, and compact, features all found in UTL.

A third possibility, easier to implement and explained with more detail in the next section, is that of using UTL in an already built interlingual MT system, as a tool to produce interlingual documents free of errors.

UTL and Machine Translation Interlinguas

In our project we are not at all the first to observe the special qualities of Esperanto as a language which can deal better with computers. There was in the 80's a long-term MT project carried out by the Dutch company BSO, where a modified version of Esperanto was used as an intermediate language between source and target natural languages (Witkam 1983). It received the name of Distributed Language Translation (DLT), and the modified Esperanto devised for serving as its interlingua represents a good starting point in the preparation of UTL, though some adaptations seem necessary given the differences in the roles of both languages. The designers of the DLT system believed that a language like Esperanto would be a better meaning representation system than a symbolic representation, as only a human language can preserve all the contents expressed by other human languages (Schubert 1992).

However, no matter how suitable an interlingua for MT may be, the automated conversion from the source language to the interlingua remains an unsurmountable hindrance. As we pointed out in the introduction to this paper, several methods have been proposed to improve the performance of the system at this stage. They include pre-edition of the source text and interaction with the machine by helping it resolve the ambiguities it finds in the process. These techniques are not free from limitations, and their application remains very scarce so far.

A Practical Implementation

The UTL project began in 1999 as a theoretical study and has been developed so far as a freelance endeavour. The possibility of having an artificial human language based on Esperanto that could express unambiguously every possible sentence was real, as the DLT project confirmed. This led directly to the ideas we have exposed above of UTL as a special kind of controlled language. Later on, in the search for a practical realization of the system, it became obvious that our language could also be used within an already built MT interlingual system. In these systems, writing directly in the interlingual code may be the most accurate way to prepare an interlingual document free of errors, but it would be a quite slow and complicate task even in hands of a trained user. The UTL language may serve here as a language-interface which allows programming interlingual code in an indirect and faster manner, in the same way that, in computer science, high-level languages are used to program in a faster and simpler way than with low-level languages.

In the last year, a small prototype based on this idea has been developed in the Computer Science Faculty of the University of A Coruña (Spain). In this sample program, UTL's concept has been adapted for the UNL system (the similarity in the name is purely casual), an interlingual MT project currently under development at the Institute of Advanced Studies of the UNU in partnership with other research institutes, universities, and R&D groups in several countries.

Our UTL application has not more than demonstrative purposes, limited to a restricted vocabulary and typology of sentences, but it performs the four basic operations envisaged for such an utility: 1. input of sentences in the UTL language; 2. analysis (and tree representation) of those sentences; 3. translation into UNL's interlingual code; 4. delivering of that code to UNL's converter to natural language.

The Problems of Machine Translation

The translation process carried out by a computer can be roughly described as having two stages in the systems using an intermediate language (interlingua). The first stage is the analysis of the source text, in order to convert it to the interlingua. Then, from the interlingua representation, the second stage is undertaken: the generation of the target language text. It is known that the most complicated problems and the mistakes which more deeply affect the final result happen during the first stage, that is, the analysis or enconversion. As we have already explained, it is at this point where the employment of UTL can make a difference, preventing the system from having to deal with problems it cannot solve by itself.

Inspired by the bibliography referring to this subject (Hutchins and Somers 1992: 81-130), we have made a short classification of the problems of analyzing and transferring natural language and we have added some examples on how UTL gets rid of them:

- Morphological problems: normally, the inclusion of a model of morphological analysis is regarded in MT systems as necessary, to reduce burden in lexicons and recognise unkown or derivate words not included in the lexicon or simply unknown to the system. However, the difficulty in dealing with irregular forms and complex paradigms damages the efficiency of this solution.

In UTL, however, morphology is extremely simple. Common processes such as conversion (from adjective to noun, from noun to adjective, etc), inflexion, and derivation are completely regular. The whole system is defined with a few rules that have no exceptions.

- Lexical ambiguity: either categorial (very common in English, where the same word can sometimes have different categories: use, control, work...); homography (two different words with a common written form: light, bank...); or transfer ambiguity (when one word in one language covers a semantic range which is covered in another language by several: English "corner" > Spanish "rincón" and "esquina").

Categorial ambiguity does not exist in UTL-Esperanto, where every part of speech has its own characteristic ending (-o for nouns, -a for adjectives, etc). In accordance with this, a same word cannot be used for two different functions; if function varies, ending changes accordingly. For example, lets compare the English word "work" with its Esperanto counterparts:

> English: *work* (v), *work* (n), *work* (adj)
> Esperanto: *labori* (v), *laboro* (n), *labora* (adj)

Homography is also excluded from our language. Every word has only one meaning in Esperanto, and the few exceptions to this have been fixed in the UTL version.

- Structural ambiguity: when there are several possible deep structures for a given sentence, in accordance with the grammatical definition used by the system. In this regard, the two possible meanings of certain English sentences are expressed distinctly in Esperanto:

> English: *Cleaning fluids can be dangerous.*
> Esperanto: 1. *Purigaj fluidaĵoj povas esti danĝeraj.*
2. *Purigi fluidaĵojn povas esti danĝere.*

Another usual case of structural ambiguity comes from the double adjunctability of prepositional phrases or adverbial phrases:

> English: *I saw a man with a telescope.*
> Esperanto-UTL: 1. *Mi vidis homon kun teleskopo.*
2. *Mi vidis homon de kun teleskopo.*
(preposition "de" is used in UTL before a PP to link it to the noun).

- Anaphora: when a word (personal pronouns, demonstratives) makes indirect reference to another entity mentioned explicitly in another place of the text.

In UTL the pronoun can be tagged when the context is confusing. For example:

> English: *The garden had a tree. I saw it.* (What did I see, the garden or the tree?)
> Esperanto-UTL:

1. *La ĝardeno havis arbon. Mi vidis ĝin (>ĝardeno).*
2. *La ĝardeno havis arbon. Mi vidis ĝin (>arbo).*

- Idioms: expressions whose meaning is not deductible from the words which form them. An idiom usually makes no sense when translated literally to another language. When writing in Esperanto-UTL, no idiomatic expressions are used. The UTL user writes being conscious that the text is going to be machine-translated.

Practical Usage

At this point of the article, the reader may think "Well, so this UTL language has been modeled in such a way that it will translate quite well to other languages, but you have to learn first the UTL language in order to use it! Isn't this too complicate?". First, we must remember that the UTL system is mainly conceived as a tool for professionals who want to make multilingual translations. This requires a prior preparation (as many other professional activities), and the UTL-Esperanto language can be learnt by anyone in a few months. It is not necessary to learn the whole language (with its thousands of words) to make a good use of the UTL system: the system holds the possibility of using English (or other natural language's) words, conveniently marked, within the UTL text. This makes possible to use the system even with a limited knowledge of the UTL language. On the other hand, this feature also solves the problem of any shortage of vocabulary that UTL-Esperanto may have in certain specialized domains.

As we said early, the project is currently being developed without any financial support, which limits the prospects of the first prototype to a demonstrative application. However, a happy fact about UTL is that it is a system fairly cheap to develop, in comparison to other MT systems, as it does not involve the computerized treatment of a complicate, irregular natural language, but a simple, artificial one. Having said all this, we just want to express our hope that the UTL system, after a complete implementation, will become a tool for translators to widen the scope of their work and to increase their productivity. Consequently, we expect that, by means of UTL, the high costs entailed by any process of multilingualization will be reduced and made affordable to a wider range of organizations, companies and individuals.

References

Boitet, C. and Blanchon, H. 1993. "Dialogue-based machine translation for monolingual authors and the LIDIA project". In Nomura, H. (ed.), *Proceedings of the 1993 Natural Language Processing Rim Symposium.* Fukuoka: Kyushu Institute of Technology, 208-222.

Hutchins W. J. and Somers H. L. 1992. *An Introduction to Machine Translation.* London: Academic Press Ltd.

Janton, Pierre. 1973. *L'espéranto* (coll. Que sais-je? no. 1511). Paris: PUF. Consulted edition: translation to Spanish by Damià de Bas, *El esperanto*. Barcelona: Oikos-tau, 1976.

Pym, D.J. 1990. "Pre-editing and the use of simplified writing for MT: An engineer's experience of operating an MT system". In Pamela Mayorcas (ed.), *Translating and the Computer 10*. London: Aslib, 80-96.

Schubert, Klaus. 1992. "Esperanto as an intermediate language for machine translation", *Computers in Translation: A Practical Appraisal.* London: Routledge, 78-95.

Witkam, A.P.M. 1983. Distributed Language Translation. Feasibility Study of a Multilingual Facility for Videotex Information Networks. Utrecht: BSO.