A MULTI-LINGUAL WEB-BASED SURVEY FORM MACHINE TRANSLATION MECHANISM

Richard Boddington [1], Judy Clayden [1], Michael Collins [1], & Sam Pride [1]

[1] School of Computer and Information Science, Edith Cowan University, Perth

ABSTRACT: Translation costs restrict the preparation of medical survey questionnaires in migrant and Aboriginal communities in Western Australia. This is further compounded by a lack of affordable and accurate machine translation (MT) mechanisms. This research investigates combined strategies intended to provide an efficacious and affordable machine translator to meet these needs by: (a) using an interlingua or hub-language which requires less resources for its construction than contemporary systems and has the additional benefit of significant error reduction; (b) creating word group structures to reduce the complexity of translation rules and enhance correct transfer of meaning between natural languages; and (c) defining smaller operating environments to restrict data input and further enhance translation by reducing error occurrence significantly. This research aims to produce a prototype MT mechanism that will accept questionnaire texts as discrete questions and suggested answers from which a respondent may select. The prototype will be designed to accept as input, non-ambiguous English as the source language (SL), translate it to a 'hub-language' or interlingua based on Esperanto and thence to a selected target language (TL). Additionally, an inverse path of translation from French back to English will enable validation of minimal or zero change in both syntax and semantics of the original input.

INTRODUCTION

This study was prompted by an urgent need to provide an affordable MT mechanism, capable of producing reliable, high-accuracy translations. The study will assist in an existing project requiring such technology: a web-based survey for Princess Margaret Hospital that seeks feedback from the parents of cancer-affected child patients. Many of these parents either do not speak English or only comprehend it in a limited way. It is important to capture their input during the surveys, but there is a high cost in terms of time, expense and availability of suitable translators for translating the surveys into a variety of natural languages. Disappointingly, some MT applications may cost tens of thousands of dollars and are often beyond the reach of many potential users (Gross, 1992). More affordable applications, including Déjà Vu; the Windows XP language translator and some of the freeware programs available on the Internet, notably the Babel Fish translator, do not provide accurate high-level translations suitable for survey research (Boitet, 1994). There are some high accuracy translation mechanisms for languages including English, French and other European languages, such as the SYSTRAN programme (Boitet, 1994), but there is a paucity of reliable systems for other languages including Aboriginal English and Aboriginal languages. It is hoped that this research will lead to the development of a MT model for use within the Australian language environment.

Detailed description of hypothesis

The shortage of suitable and affordable applications may be attributable to the challenges of developing MT programs compared to other applications such as electronic dictionaries, spelling and grammar checkers. The additional resources and time spent in developing MT applications significantly increases the purchase and licensing costs (Boitet, 1994). Preliminary research has examined these challenges and has focussed on mechanisms offering solutions worthy of formal feasibility testing. These solutions include the use of an interlingua or hub-language; the formation of word group identification to enhance translation fluency; and the restriction of the size of language dictionaries to reduce syntax and semantic translation rules.

An interlingua may overcome some of the shortcomings prevalent in traditional automated translation mechanisms. Traditional systems have several limitations because (a) they use a dictionary customised for each language pair (Boitet, 1994); and (b) often use poorly structured syntax and,

consequently, produce erratic, inconsistent results (Gross, 1992). The proposed study aims to counter (a) by using an established, strictly rule-based interlingua as a means of translation, rather than creating a language paired translation mechanism. The use of language pairs necessitates the production of n(n-1) mechanisms (where 'n' is the number of selected languages) to achieve translation (Kay, 1996), whereas an interlingua requires only 2n translation mechanisms to achieve the same results (Arnold, Balkan, Meijer, Lee Humphreys, & Sadler, pp. 74-75). It also aims to offset (b) above by restricting translation input within defined syntactic and lexical boundaries and by using word grouping analysis rather than individual words to preserve meaning and provide natural expression in the TL.

For an interlingua to fulfil its role as an accurate translation template it must express ideas in a way similar to most major natural languages and therefore be capable of expressing meaning in a precise and unambiguous manner (Sabaris, Alonso, Dafonte, & Arcay, 2001). To create a new interlingua would be a laborious and unnecessary venture, as Esperanto, according to Sabaris et al., is

> . . . seen 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 [An experiment using Esperanto as a universal translation mechanism] 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 (Sabaris et al., 2001).

A small MT model will evaluate the accuracy of translation between two natural languages: namely, English and French. A software program built around the relationship between the natural languages and the interlingua will facilitate the accurate transfer of meaning between the natural languages. Evaluation of translated texts by qualified linguists in English, French and Esperanto will determine the level of translation accuracy between these entities. Successful completion of this research will realise a step towards an increased availability of low cost MT to assist in the development of reliable and efficient survey translation systems for use in specific user environments. It is hoped that the research will lead to the development of a MT model for use within the Australian indigenous language environment. Another incentive acknowledges research demonstrating high-accuracy translation results with MT systems relying on smaller, specialised vocabularies and producing significantly high quality translations (Arnold et al., 1994, pp. 7, 150-151). Limiting the size of the syntax and grammar of the languages used in the prototype should guarantee a commensurate error reduction, without loss of translation capability, in each survey, while the development of word group identification enhances semantic translation (Henisz-Dostert, Macdonald, & Zarechnak, 1979, pp. 15-16).

THE SIGNIFICANCE OF THE STUDY

Encouraging published research results suggest that:

(a) An interlingua model requires less resources for its construction than contemporary systems (Kay, 1996) and has the additional benefit of significant error reduction in preserving semantic meaning of the original text (Arnold et al., 1994; Schubert, 1988, 1997, 1998; Witkam, 1988);

(b) Development of word group identification enhances semantic translation (Henisz-Dostert et al., 1979, pp. 15-16); and

(c) Smaller operating environments with formal input enhance translation by reducing error occurrence significantly (Arnold et al., 1994, pp. 150-151).

As a result, there is the potential to achieve high-level translation and greater simplicity in model design and construction. It should be possible to harness these features into a prototype mechanism that will effectively translate targeted phrase styles, i.e. those found in well-designed surveys. Likely benefits will include simplified and less expensive translations of surveys, initially for the academic community. However, a successful prototype may well have commercial applications and may attract continued funding from survey vendors.

THE PURPOSE OF THE STUDY

The purpose of this study is to: (a) investigate the feasibility of constructing a prototype MT mechanism using Esperanto as an interlingua: (i) within the scope of small budget environment; (ii) contained within a defined range of formal natural language; (iii) incorporating a word group identification process; and (iv) whose purpose is to generate high-level translation output; and (b) determine the feasibility of using an interlingua based on Esperanto to satisfy the needs of survey designers. The anticipated outcome of the project will be a pilot implementation of an interlingua MT mechanism; using formal natural languages and incorporating defined lexicons and syntaxes. Evaluation of the output translations will determine the level of translation accuracy including semantic accuracy output. The project will also examine the efficacy of the lexicon and syntax structures contained in the word group modules. Analysis of the relationships between the groups that make up each sample text will assist in the design of improved translation rules for future models. In addition, these evaluations will determine whether a reliable and inexpensive mechanism may be created for survey design.

STATEMENT OF RESEARCH QUESTIONS

This research has been designed to examine the overall research question:

> May an Esperanto-based interlingua MT mechanism be used in a defined natural language environment with a word group identification process to achieve high-accuracy translations at low cost between two natural languages?

REVIEW OF RELEVANT LITERATURE

'Indirect' or 'linguistic knowledge' (LK) architecture dominated MT research in the 1980s, taking the lead from its predecessor, 'transformer' architecture. LK architecture places considerably more reliance on linguistics, specifically relying on a detailed understanding of both the SL and the TL (Gross, 1992). Arnold et al. (1994, pp. 66-69) point out that the translation from the SL to the TL is intended, but in practice most programs have problems ensuring that the lexical rules work in both directions. These transfer-based designs influenced the development of new approaches and deeper research into an interlingua based MT. Some MT developers recognised that the deeper they investigated disassembling natural languages to translate the original meaning and then transfer it with full integrity into the SL, the more some form of abstract intermediary was required. This intermediary stage in which the meaning is common to both the SL and TL, may be described as an interlingua system (Arnold et al., 1994, pp. 75-76; Witkan, 1988).

Carnegie Mellon University, Pittsburgh, carried out research into knowledge-based systems within the Artificial Intelligence (AI) community. Arnold et al., (1994, p. 77) expand this research into the use of an AI based interlingua:

> The argument is that MT must go beyond purely linguistic information (syntax and semantics); translation involves "understanding" the content of text and must refer to knowledge of the "real world." Such an approach implies translation via intermediate representations based on (extra-linguistic) "universal" elements" (Arnold et al., 1994, p. 77).

In the 1980s, MT programs used in commercial contexts produced some improved outcomes. A diesel engine manufacturer, Perkins Engines, claimed it had made significant savings in translating engine manuals through using a version of the WEIDNER MT program (Arnold et al., 1994, p. 7). These researchers concluded that working with a smaller defined lexicon produced higher translation accuracy than if they had used larger, encyclopaedic-type dictionaries. Since 1977, the Canadian Meteorological Centre in Montreal has used METEO, an MT system capable of providing translations from English to French. Still in use in the 1990s, the program required approximately 4% human intervention to ensure accuracy of interpretation (Arnold et al., 1994, p. 7; Boitet, 1994). The significance of this program was the use of a 'sub-language' designed for communicating between experts in such areas as science, medicine or technology (Arnold et al., 1994, p. 150-151). Many of the more successful MT programs are of this nature, relying on smaller, specialised vocabularies (Arnold et al., 1994, p. 150-151; Kay, 1996).

SPECIFIC STUDIES SIMILAR TO THE CURRENT STUDY

Dutch researchers have undertaken interlingua research, using natural or artificial languages: the Distributed Language Translation (DLT) is based on a modification of Esperanto and the Rosetta system experimenting with Montague semantics (Arnold et al., 1994; Witkam, 1988). The DLT experiment, carried out by Buro voor Systeemontwikkeling in Holland, received some publicity before it ended in 1988 (Schubert, 1988). In 1999, the Universal Translation Language (UTL) proposed a new approach to multilingualization, based on the usage of an artificial unambiguous human language or interlingua (Sabaris et al., 2001). The UTL aimed to provide a tool to convert a specific text written in a given natural language into an indefinite number of other languages in a process of human assisted MT:

> 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 (Sabaris et al., 2001).

According to the project developers, a small prototype was developed in the Computer Science Faculty of the University of A Coruña in Spain. In this sample program, the concept was adapted for an 'interlingual' MT project currently under development at the Institute of Advanced Studies of the UNU [sic] in partnership with other research institutes, universities, and research and development groups in several countries. Details of the efficacy of the UNU prototype were not available at the time this proposal was developed.

MATERIALS AND METHODS

The experiment will involve the use of IBM computers and the "Jade" object-oriented environment to create the MT software and its underlying database. The translation instruments will include reliable dictionaries and resources capable of ensuring accurate translations and syntax between the three languages used in the experiment. Checklists will be used to record the collected data. These checklists will validate the spelling, syntax and interpretation of the sample texts tested during the evaluation process.

Recording will be made of texts adapted from the Princess Margaret Hospital survey document, in the form of word groups comprising one or two clauses selected from the SL. The texts entered into the prototype mechanism will then undergo a process of translation through the interlingua, which will convert the data into reciprocal texts in the TL. Reverse processing, using the output texts in the TL for translation through the interlingua to the SL, will then enable comparison of the output result compared with the original sample text. Written recordings of the data samples tested will enable identification of errors at each stage of the process through translation verification. This will facilitate overall efficacy of the prototype model. The comparison of the data relating to the resources and costs involved in the design, construction and testing of the prototype with those of a hypothetical, non-defined, non-interlingua model requires separate collection procedures. At each key stage during the project, data collection will take the form of recording human and technical resources involved. Time spent on individual tasks, costs and levels of expertise required to complete these tasks will be tabulated and will form a basis for data analysis. A similar process to evaluate the hypothetical model will use the same collection method.

Examination of the collected translation data relies on the accuracy of translation from the SL to the TL and further comparison between the re-transfer of that data from the TL to the SL. Crucial to the proposed thesis is the occurrence of errors at both stages of the operation. A zero-error occurrence is the desired outcome. However, there is an expectation that errors of incorrect syntax and word or word-group definition may occur. Such occurrences do not necessarily detract from the efficacy of the prototype, but will require analysis to investigate whether the causes are a construction fault in the syntax and/or word-grouping design and whether or not they may be resolved. Comparison between the resources and costs used in the construction of the prototype and conventional methods will determine whether there are significant differences between the two methods.

To produce a working model capable of meeting the requirements of a survey user requires greater time, resources and budget than available to the authors. These factors will limit the size of the proposed prototype in terms of syntax and dictionaries. However, the proposed model will provide the opportunity to test a sufficient amount of sample data to determine whether larger-scale testing might be viable.

REFERENCES

Arnold, D., Balkan, L., Meijer, S., Lee Humphreys, R., & Sadler, L. (1994). *Machine translation: an introductory guide*. Oxford: NCC Blackwell Ltd.

Boitet, C. (1994). *(Human-aided) machine translation: a better future?* Grenoble, France: Université Joseph Fourier.

Gross, A. (1992). Limitations of computers as translation tools. In J. Newton (Ed.), *Computers in Translation: A Practical Appraisal*. London: Routledge.

Henisz-Dostert, B., Macdonald, R. R., & Zarechnak, M. (1979). *Trends in linguistics: studies and monographs 11: machine translation*. The Hague: Mouton Publishers.

Kay, M. (1996). *Machine translation: the disappointing past and present*. Palo Alto, California, USA: Xerox Palo Alto Research Center.

Sabaris, M. F., Alonso, J. L. R., Dafonte, C., & Arcay, B. (2001). *Multilingual authoring through an artificial language.* Paper presented at the MT Summit VIII, Santiago de Compostela, Spain.

Schubert, K. (1988). The architecture of DLT: interlingual or double direct. In *New directions in machine translation*. Dordrecht, Holland: Floris Publications.

Schubert, K. (1997). *DLT: The facts*. Retrieved 23/02/02, from http://www.sudakruco.org/032/english.html

Schubert, K. (1998). The architecture of DLT: interlingual or double direct? *In Maxwell et al. [1988], pages 131--144.*

Witkam, T. (1988). *DLT: An industrial R & D project for multilingual machine translation.* Paper presented at the Proceedings of the 12th International Conference on Computational Linguistics, Budapest.