Machine Translation and Multilingual Communication on Internet

M. Abdus Salam
Faculty of Informatics and Communication
Central Queensland University
Rockhampton, Australia 4702
e-mail:a.salam@cqu.edu.au

Abstract

The best cognitive model of a multilingual system is drawn from the corpora of multilingual communication on the internet and the human experiences of the people with multicultural and multilingual backgrounds. Such a system provides the basis of a multicultural context and knowledge that can be used for an effective machine translation.

Keywords: Context, Multilingual Translation, Internet

1 Introduction

The internet Asian communities normally use English script to express their own languages along with English. It gave rise to electronic dictionaries to standardise these practices as on web dictionaries. Interestingly the internet communication quite often involves multiple languages particularly those who live in the countries where they speak second and/or third language tend to mix a number of languages. Many *soc.culture* newsgroup messages reflect this trend. Languages are mixed in a number of ways; either entire sentence or phrases are used or a particular word is inserted in a sentence. This is a reflection of the fact that an accurate translation is an illusion. People use word and phrases from other languages because what they want to convey is better expressed in another language. But more importantly these practices lead to a very different approach to translation. In the process of mixing the languages a very interesting phenomenon is occurring, that is, people are building a context based that is multicultural in nature and hence multilingual in structure.

We develop ideas and concepts in the context of culture, society, time and space where we experience life. These contexts assign certain meanings to the language we use. In that sense the context determines what a certain expression means in a particular time and space. The more we know about the context the better understanding we will have. This approach has been taken in the current research in Natural Language Processing and machine translation and may be found in Akman (1997), Buvac (1996), Hausser (1999), McCarthy (1997) and Salam (1999a). While it is possible to manipulate text in a particular language for information retrieval using context, it is not clear how this can be done when we change the language as well as is the case in machine translation. An attempt has been made in Salam (1999b). Some of the earlier developments in machine translation of particularly Asian languages can be found in Mitkov (1993), Yusoff (1992) and Yusoff (1993).

In this paper first we give the background of context based systems and approach to machine translation, then show how the development of multilingual communication can help us in developing a theory of language which explains the natural transfer of information in a way that is functionally coherent, mathematically explicit and computationally efficient.

2 Context Based Systems

The formalisation of context is not interested in solving the syntactic and semantic problems. Such a formalism requires a well defined context when dealing with discourse. A *discourse context* has to be characterised by the intended meaning of their predicates.

The basic relations used in formalising contexts as *objects* are:

- *ist*(c,p) means the proposition p is true in the context c and
- value(c,e) assigns the value of the term e in the context c.
- We also have an *existence* predicate that determines whether an object x exists or not in a given context c. It is denoted as $E(c,x)$ and is true if, and only if the object x exists in the context c.
- The above relations are single valued whereas the relation *meaning*(c,w) introduced in Salam (1999b) is a list of all the "meanings" of the word or phrase w in context c.

In a recent work, Bonzon (1997) has given a reflective proof system for reasoning in contexts. Using a language of contextual implications, Bonzon develops a system of natural deductions of contextual facts.

Our objective is to process a given piece of prose that may contain either factual information or fictional propositions including the ones that involve more than one *values* or *meanings* in different contexts that may exist at the same time. As pointed out in McCarthy (1997), in natural language, context may vary on a very small scale and several contexts may occur in a single sentence. In most cases important contexts are implicit and imbedded in the text itself. As a result we have outer contexts that are generally known and inner or implicit contexts that can only be extracted from the text itself. The understanding of these implicit contexts is essential in processing a natural language for extracting the information needed for a proper translation.

2.1 Context as Knowledge Base

What McCarthy and Buvac describe as *rich contexts* are actually contexts depending on a knowledge base of a human or a machine which can never be assumed complete. For example the truth value of *ist*(c, p) where c is the context of Australian history and p is the statement, "Malcolm Fraser was a prime minister." depends on the interpreting object's knowledge base of Australian history. Hence each context is dependent on a knowledge base variable. Such a situation has not been much discussed in literature but is crucial in understanding natural language. We also consider time as in space-time coordinate system, a context that may be explicit in the text or implicit in form of grammatical structure of the sentences.

3 A Basic Translator

As a human translator we know that the single most important criterion for an effective translation is the knowledge of the languages and the contexts in which the particular piece of discourse was produced. A typical translator will require:

- a multilingual parser,
- a knowledge base context structure,
- four dictionaries and
- language generators.

It should also be capable of

- recognising and correcting typographical errors,

- extracting information from the grammatical structure of the sentences
- particularly information regarding time context.
- ignoring the grammatical errors.

For the sake of simplicity we consider a bilingual system of languages say A and B. Let C^A be the cultural context of the language A and C^B the cultural context of B. Let L_A and L^B be the lexicons of A and B then we define

Definition 1

The meaning of a word w in a language A is a mapping

$$M : A \rightarrow A$$

such that

$$M(w) = L_A^w$$

where L_A^w is the sub-lexicon of A associated with w and is the "meaning" of w in A.

We also define translation as:

Definition 2

The translation from A to B is a mapping

$$T : A \rightarrow B$$

such that

$$T(w) = L_B^w$$

where $w \in L_A$ and L_B^w is the sub-lexicon of B associated with w and is the "translation" of $w \in B$.

Given these two definitions and our discussion we propose that:

Proposition 1

The union of all sub-lexicons obtained through all the iterations of Definitions 1 & 2 for a particular word or phrase w is closed in a given context.

This union of sub-lexicons is generally large but in order to reduce it to the minimum we need as much knowledge of the context as possible. To start with we combine the two cultural contexts namely C_A and C_B. This marriage of two cultural contexts is based on the vast multilingual and multicultural corpus available on the internet. Let us call it C_{AB}. At this stage we apply the relation *meaning*(c, w) to obtain the list of possible words and phrases in the translation. The idea is to apply this relation at every iteration.

This whole process is further enhanced by the use of the *learning algorithm* from the internal context of the multilingual corpus as proposed by Edmonds (1999).

5 An Example

Consider the following conversation in a bilingual environment where English and Urdu are mixed together.

Salman: *yaar, tum ney kuch suna ye tariq aur margaret ka qissah*?
Rushdie: *I know, range hathoN pakrey gaey.*
Salman: *bahar sey pata nahiN chalta tha key there was any connection.*
Rushdie: *arey budhoo, they knew each other and I mean in Biblical sense.*
Salman: *I see!*

It is obvious these people know both the languages and the cultural contexts. How do I know it? Well, I have spent virtually equal amount of my life in both the cultures and know the cultural roots of the respective languages. It is the cognitive model of such natural multilingual systems that facilitates any meaningful translation. The dictionaries I most rely on for hands-on translations are not the ones published but the ones I have developed over the years and are part of my knowledgebase. These dictionaries are context dependent and evolving through the learning algorithms Edmonds (1999).

Let us see how it works. First of all a bilingual parser should be able to parse Urdu and English parts. The next step would be the identification of the context structure.

Translation to English: If we decide to translate the above conversation to English, we look at the Urdu part in a multicultural context. The internal context is set by the opening sentence that it is about a man called Tariq and a woman called Margaret. Some parts are relatively easy as is the case with the first sentence. Let us have a look at some difficult parts of the conversation. Consider, "range hathoN pakrey gaey". Looking up in any dictionary will not give any meaning for "range" because "range hathoN" is a complete phrase which literally means, coloured hands. An idiomatic translation will give *red handed*. And "pakrey gaey" is easily translated to *got caught*. Now the internal context is made further clear that the two people were up to something together. This helps translation of the next bit that involves the phrase, "pata nahiN chalta". Literally, $L^{pata}_{English}$ includes {sign, symptom, clue, address to which a letter is directed, hint, token} Feroz sons. In this context there is no letter involved explicitly or implicitly then the address is out of question. The rest may still apply. Combining the other three words we get a phrase, "pata nahiN chalta tha" that in this context means, *had no sign* or *had no clue or hint*.

Translation to Urdu: The most difficult part to translate to Urdu for Urdu speaking readers that has no or little knowledge of the socio-religious background of English speaking cultures will be the last sentence. The second part of the English sentence, "I mean in Biblical sense" sets the internal context of the preceding part, "they knew each other".

A literal translation will mean nothing to a Urdu speaking person unless knows the English version of the Bible which is not common.

6 Concluding Remarks

We feel that a multilingual context and knowledgebase is essential in the modelling of any cognitive system for translation. Such a model is drawn from human experiences with the capacity to evolve and change with time and space. we have demonstrated that such a system is possible and workable.

References

Akman V. and Mehmet S. (1996), 'Steps toward formalizing context', in AI Magazine 17(3) 55-72.

Akman V. (1997), 'Rethinking context as a social construct', in AAAI-97 Fall Symposium on Context in Knowledge Representation and Natural Language, MIT Cambridge.

Bonzon P. (1997), 'A reflective proof system for reasoning in contexts', in Proceedings of 14th National Conference on Artificial Intelligence, AAAI.

Buvac S. (1996), 'Quantificational logic of context', in Proceedings of 13th National Conference on Artificial Intelligence, AAAI.

Buvac S., Buvac V. and Mason I. (1995), 'Metamathematics of context', in Fundamenta Informaticae 23 3.

Edmonds B. (1999), 'The Pragmatic Roots of Context' in Bouquet, P. et al. (Eds.), Modeling and Using Context: the Proceedings of CONTEXT'99, Trento, Italy. Lecture Notes in Artificial Intelligence, 1688:119-132.

Hausser R. (1999), Foundations of Computational Linguistics, Springer:Berlin.

McCarthy J. and Buvac S. (1997), 'Formalizing Context' in Computing Natural Language, Editors Aliseda et.al, Stanford University.

Mitkov R. (1993), 'Sublanguage schemata and their importance in machine translation', in Proceedings of the International Conference on Mathematical Linguistics, Tarragona.

Salam M. A. (1999a), 'Context Based Natural Language Processing' in Proceedings of the International Conference on Information, Communication and Signal Processing, Singapore.

Salam M. A. (1999b), 'A Step Towards Context Based Machine Translation of Asian Languages', in Proceedings of MAL'99, China.

Urdu-english Dictionary, Ferozsons (Pty.) Ltd. Karachi. ISBN 969 0 00508 1.

Yusoff Z. (1993), 'The role of grammar formalisms in machine translation', in Summer School: Contemporary Topics in Computational Linguistics, Bulgaria.

Yusoff Z. and Lepage Y. (1992), 'On the specifications of abstract linguistic structures in formalisms for machine translation', in International Symposium on Natural Language Understanding and Artificial Intelligence, Kyushu.

A Web of On-line Dictionaries, http://www.yourdictionary.com/