An Interactive Approach to Development of English-Tamil Machine Translation System on the Web.

Dr. Vasu Renganathan
University of Pennsylvania, PA, USA

0. Introduction

One of the potential applications of Natural Language Processing (NLP) research, in general, is believed to be development of Machine Translation (MT) applications, because a successful and all purpose machine translation system obviously would save a vast amount of human energy and time for translation of one language text to another. There are MT systems which usually produce output that requires much of post-editing by human. "... despite extensive research in this area, which started in 1950s, little progress has been made. The reasons for the lack of success are linguistic and computational. On the one hand, linguistic theory does not provide answers to all problems of machine translation; on the other hand, software and hardware problems restrict the implementation and design of machine translation systems to a certain degree" (Handke, 1995:11). However, producing robust systems requiring only a limited amount of human intervention on the output would still be a major task in this field of research.

1. Scope of the system

This paper illustrates a research being pursued on development of English-Tamil machine translation system. A work-in-progress version of this system may be tested online in the URL: <http://lrrc3.plc.upenn.edu/tamil/.> This is a rule-based system containing around five thousand words in lexicon, and a wide range of transfer rules written in Prolog encompassing frequently occurring English structures mapped to corresponding Tamil structures. Both rule base and the lexicon of this system are built in such a way that the users can update the scope of this system interactively by adding words into lexicon and rules into rule-base. Translating both colloquial and technical English into Tamil with a computer essentially involves construction of the two basic blocks namely the lexicon and rule base. Construction of online lexicon requires codification of grammatical information in two different ways. One by coding a minimal-set of information about grammatical categories of head and target words and the other by including an extensive information involving semantic and syntactic properties of words. The former type of lexicon is sufficient for translating technical, colloquial and news documents, where as the latter type of lexicon is mandatory for translating complex type of literary texts comprising fiction, poems, biographies etc. The system demonstrated here is built with the former type of lexicon containing a minimal set of grammatical information for head words of both English and Tamil. The other significant component of any machine translation system is building a rule base that maps the structures of both source and target language. Any ideal system should be capable of accommodating not only the basic structures of source language, but also a wide variety of complex structures accounting for all sort of ambiguous interpretations. The programming language Prolog facilitates a logical method for codifying such complex rules in a robust way.

The aim of this paper is not only to sketch the algorithm employed in this system, but also to state the feasibilities of further research in this area. Problems of translation, in general, range from lexical, linguistic and discourse phenomena. While accounting for such multifaceted problems is still a daunting task for human translators, it might even be an unsolvable question in the context of machine translating such complex texts. However, exploiting the commonness across languages and setting the right parameters to transfer text from source language text to target language text would not be too difficult to accomplish. Further, transferring exactly the contextual, metaphoric and figurative expressions employed in the source language text to the target language text would produce substantial result. Accounting for infinite number of rules of any language in any translation system is almost an impossible task. However, one can always customize the MT systems in such a way that they can be updated to accommodate new rules, without much effort. For example, using the updatable external database entries namely vbdb.dat, engvb.dat, and lexicon.dat that are used as part of this system, it is possible to let this system translate text consisting of new rules and lexical entries. Consider below the translation rendered by this system of the poem "The Gardener" by Rabindranath Tagore into Tamil.

2. Example Input and Output by the MT system:

Gardener. தோட் டக்காரர்.
rabindranath tagore ரபீந்திரநாத் தாகூர்.

if you would have it so, i will end my singing
நீங்கள் இதை அப்படி வைத்திருந்தால் நான் என்னுடைய பாடுவதை நிறுத்துவேன்.

if it sets your heart aflutter, i will take away my eyes from your face
இது உங்களுடைய இதயத்தை படபடக்க செய்தால் நான் என்னுடைய கண்களை
உங்களுடைய முகத்திலிரு ந்து எடுத்துவிடுவேன்.

if it suddenly startles you in your walk, i will step aside and take another path
இது உங்களை உங்களுடைய நடையில் திடீரென்று தடுமாறச்செய்தால் நான்
நகர்ந்துபோய் இன்னொரு வழியை எடுத்துவிடுகின்றேன்.

if it confuses you in your flower_weaving, i will shun your lonely garden
இது உங்களை உங்களுடைய பூப்பறிப்பதில் குழப்பினால் நான் உங்களுடைய தனியான
தோட் டத்தை தவிர்த்துவிடுவேன்.

if it makes the water wanton and wild, i will not row my boat by your bank
இது தண்ணீர் குழப்பத்தும் கலக்கத்துமை செய்தால் நான் என்னுடைய படகை
உங்களுடைய கறைக்கு பக்கத்தில் துடுப்புப்போடமாட் டேன்.

Despite the fact that this target text in Tamil is not as idiomatic as one would find in the case of its equivalent English version, the syntactic transfer that is made use of in this system is capable of retaining the metaphoric and figurative expressions from the source text. Technical documents, as opposed to fictions, are generally devoid of such metaphoric and figurative usages, and thus are less problematic to handle by the system like this one. Consider below a sample translation of technical text.

Input: This is the external appearance of a normal heart. The epicardial surface is smooth and glistening. The amount of epicardial fat is usual. The left anterior descending coronary artery extends down from the aortic root to the apex.

Output: இது ஒரு சாதாரண இதயத்தினுடைய வெளிப்புற தோற்றம். உள்பக்க மேல்பகுதி நெகிழ்வானதும் பளபளப்பானதும். உள்பக்க கொழுப்பினுடைய அளவு சாதாரணமானது. இடதுபுற பின்பக்க இறங்குமுகமான விளிம்பக தசைக்குழாய் கீழ்முகமாக மையத்தண்டு வேரிலிருந்து உச்சிப்பகுதிக்கு நீள்கின்றது.

The post-editing part of this translation not only requires making necessary changes in incorrect word forms like *குழப்பத்தும், *கலக்கத்துமை as above, and also requires possible modifications in syntax and selection of lexical equivalents. Although, one does not have much freedom to decide upon the transfer rules built as part of this system in a top-down parsing strategy, making necessary changes in lexicon to alter the overall translation is still a possibility.

2.1.. Lexicon

Building a robust lexicon and implementing an extensive search utility are the two significant aspects of any effort toward making applications for machine translation system. The example entries below from the lexicon of this system show how different part-of-speech categories are handled by this system within a specific number of fields in each record.

The seven way classification of Tamil verbs with pertinent subclasses reflecting all Tamil morphological inflections of verbs on one to one basis is employed as part of the lexical entries for verbs. The morphological transducer built as part of this system uses this information to generate correct inflectional forms. The Person, Number and Gender markers such as “atu”, “ana”, “aar”, “aarkaL” etc., are marked as part of the entries for nouns. This information is used by the syntactic parser to determine the concord relationship between the subject and verb in output sentence. Except for the irregular English verbs, all the regular verbs are marked in their present tense form, so the English morphological analyzer can parse English word forms into their root,. Cases of adjectives adverbs etc., as in the case of the entries like ‘suddenly’, ‘lonely’ as noted above, are not however, accounted for by a processor yet. An ideal system would account for all such word formations of both English and Tamil, so the size of lexicon can be manageable effortlessly. Further, compound words are dealt with in this system in two ways with or without a delimiter as in the case of word("flower_weaving", "noun","puuppaRippatu","atu"). The other possibility is to let the system make a compound itself and verify the validity from the entries stored in a separate database of compound nouns and verbs. The syntactic parser and transfer strategies are handled by a set of Prolog modules such as cp(), s(), npmax(), np(), vpmax(), vp() etc. These modules are meant for both parsing the constituents from input English sentences and also to make the transfer of structures to Tamil.

word("aflutter","pr","paTapaTa","6")	word("lonely","adj","taniyaana","")
word("another","adj","veeRoru","")	word("make","pr","cey","1")
word("aside","adv","","")	word(“came”,”pas”,”vaa”,”2d”)
word("away","adv","","")	word("path","noun","vazhi","atu")
word("bank","noun","kaRai","atu")	word("relation","noun","toTarpu","atu")
word("boat","noun","paTaku","atu")	word("row","pr","tuTuppuppooTu","4")
word("confuse","pr","kuzhappu","3")	word("set","pr","cey","1")
word("direct","adj","ndeer","")	word("shun","pr","tavirttuviTu","4")
word("end","pr","ndiRuttu","3")	word("so","adv","appaTi","")
word("eye","noun","kaN","ana")	word("startle","pr","taTumaaRaccey","1")
word("face","noun","mukam","atu")	word("step","pr","ndakarndtupoo","3b")
word("garden","noun","tooTTam","atu")	word("suddenly","adv","tiTiirenRu","")
word("gardener","noun","tooTTakkaarar","aar")	word("take","pr","eTuttuviTu","4")
word("heart","noun","itayam","atu")	word("walk","noun","ndaTai","atu")
word("wanton","noun","kuzhappam","atu")	word("surface","noun","meelpakuti","atu").
word("water","noun","taNNiir","atu")	word("smooth","noun","ndekizhvaanatu","atu").
word("wild","noun","kalakkam","atu")	word("amount","noun","aLavu","atu").
word("have","pr","vaittiru","7")	word("fat","noun","kozhuppu","atu").
word("singing","noun","paaTuvatu","atu")	word("left","adj","iTatupuRa","").
word("external","adj","veLippuRa","").	word("anterior","adj","pinpakka","").
word("appearance","noun","tooRRam","atu").	word("coronary","adj","viLimpaka","")
word("normal","adj","caataaraNa","").	word("down","adv","kiizhmukamaaka","").

word("heart","noun","itayam","atu").	word("down","adv","kiizhmukamaaka","").
word("epicardial","adj","uLpakka","").	word("root","noun","veer","atu").
word("artery","noun","tacaikkuzhaay","atu").	word("apex","noun","uccippakuti","atu").
word("artery","noun","tacaikkuzhaay","atu").	word("glistening","noun","paLapaLappaanatu","atu").
word("extend","pr","ndiiL","1").	word("usual","noun","caataaraNamaanatu","atu").
word("down","adv","kiizhmukamaaka","").	word("descending","adj","iRangkumukamaana","").
word("this","noun","itu","atu").	

3. Theoretical background

The syntactic modules of this system responsible for structural transfers from English to Tamil are constructed following the concepts of the theory of Government and Binding theory, which analyzes natural language sentences in terms of a number of language independent syntactic modules (Cf. Chomsky, 1982). The morphological transducer, responsible for generating Tamil word forms, on the other hand, is constructed following the concepts of the theory of lexical phonology, which accounts for the interrelationship between phonological and morphological rules in terms of lexical and post lexical rules (Cf. Mohanan, 1986). Different types of English and Tamil sentences are accounted for by appropriate number of Prolog modules. In other words, more than one instance of cp(), as part of the rule base, would account for different complement structures such as complement clause, conditional clause, interrogative sentences and so on. Since this system is built upon translating sentences on one to one basis, and no strategy, whatsoever, is implemented yet to compare constituents across sentences, inter-sentential properties like anaphor resolution, pronoun references etc. In this sense, the scope of this system is limited in certain respects. However, as most of the

components of this system are modular in nature, they can easily be adapted and modified to build any large scale system with a wider scope.

4. Syntactic structures accounted for by this system

Following set of rules identifies some of the basic structures employed in the syntactic parser of this system. These structures and their corresponding transfer rules form the basis of this system to work around with English sentences at the base level constituents. Complex sentences are constructed by a combination of these basic blocks in a number of different ways.

i) S = NP + (PP)n + VP - Simple sentence with a subject, a verb phrase and a number of Prepositional Phrases (PP)

ii) NP = (adj)n + N - Noun phrase consisting of a number of adjectives and a noun

iii) NP = NP Wh S - Relative clause sentences with a Wh operator

iv) PP = prep. + NP - Prepositional phrase consisting of a preposition and a noun phrase

v) VP = V + adv - Verb phrase consisting of a verb and an adverb

vi) S1 + that + S2 - Complementation sentences with the complementizer that

vii) if S1 S2 - Sentences with conditional clause

5. Online MT system and significance of using Morphological Tagger

One of the advantages of using any MT system online is that it allows one to train the system with already existing and constantly growing online documents both in source and target languages. To site an example, one can make use of English and Tamil news pages containing identical news items, and attempt to build both an online lexicon and a correlative syntactic transfer rules based on such content. This type of research, however, would require an efficient morphological transducer both for English and Tamil, so the uninflected words can be extracted from source text. As part of this research and as an ongoing work, a morphological tagger with an adequate number of rules of Tamil morphology - housed under: HYPERLINK http://lrrc3.plc.upenn.edu/tamil/tagger.html/ http://lrrc3.plc.upenn.edu/tamil/tagger.html/ - is being used to produce annotated tagged text. This morphological transducer in combination with the syntactic knowledge base, which is part of the MT system, can be used to correlate texts from both English and Tamil. The present system, to some extent, can compare the rules employed as part of the input text with that of the already existing rules from rule base and produce output that can later be updated with precise transfer rules.

5..1 Input and output from the morphological transducer:

Input text: ndaaRkaaliyil oru yoocanaiyum illaamal aRaikku veLiyee uTkaarndtirundta cundtarattiRku andtac cattam talaiyil aTittatu pool keeTTatu. ezhundtu cenRu katavait

taTTinaan; oru patilum illai. (Source: Akilan's "Kanga Snanam")

Tagged Output:
[["loc","ndaaRkaali","noun"],["adj","oru"],
["nom","yoocanai","noun","conj"],["neg_avp","il"],["dat","aRai","noun"],
["nom","veLi","noun","empe"],["pa_ajp_perf","uTkaar"], ["dat","cundtaram","noun"],
["adj","andta"],["nom","cattam"],["loc","talai","noun"],
["pa","aTi","neut.sg"],["lk","pool"],["pa","keeL","neut.sg"],["period","."],
["avp","ezhu"],["avp","cel"],["acc","katavu","noun"],["pa","taTTu","3mas.sg"],
["semicolan",";"],["adj","oru"],["nom","patil","noun","conj"],["nom","illai"], ["period","."]]

The significant aspect of producing this kind of tagged corpus is that one can easily implement a translation algorithm to translate from Tamil to English or Tamil to any other language based on this tagged output of Tamil sentences, provided suitable parser and transfer rules are built accordingly. This tagger is trained to identify words and suffixes only from literary Tamil words, and no attempt is made yet to process any Tamil text containing words in spoken form. This is especially due to intricate and obscure nature of dialectal forms. As long as one restricts both the domain of the language and its variety, mechanization of the Tamil language is quite possible in many respects.

6. References

Chomsky, Noam. (1982). Lectures on Government and Binding, Foris.

Arden, A. H. (1954). A Progressive Grammar of Common Tamil. Madras: The Christian Literature Society (5th edition).

Clocksin, W.F. and C.S. Mellish (1987). Programming in Prolog. (3rd ed.)} Springer-Verlag: New York.

Cruse, D.A., (1986). Lexical Semantics. Cambridge: Cambridge University Press.

Handke, Jurgen. (1995). "The Structure of the Lexicon. Human versus Machine". Natural Language Processing. Mouton de Gruyter: Berlin. New York.

Hutchins, W. J (1986) Machine Translation: Past, Present, Future. Chichester (UK): Ellis Horwood.

Mohanan, K.P. (1986). The Theory of Lexical Phonology. Dordrecht: Reidel.
For a brief history of machine translation see Hutchins(1986). English-Tamil Machine Translation System