Prospects for Machine Translation of the Tamil Language

Fredric C. Gey <gey@ucdata.berkeley.edu>
University of California, Berkeley

Abstract

Developing commercially viable machine translation software from one language to another is a very expensive endeavor, both in terms of the software development effort required and in terms of the linguistic resources which need to be assembled. To properly implement machine translation, it is generally accepted that one needs a bi-lingual dictionary of at least 250,000 words as the basic foundation, as well as *general morphological software* for both source language and target language which will automatically parse sentences into their constituent adjective-noun-verb-object structure. Finally one needs a *transfer grammar* which maps the grammatical structure from the first language into the translated language. (For more information about the details of machine translation, the reader is referred to the excellent book **Translation Engines,** by Arthur Trujillo, published by Springer, 1999). From the point of view of the Tamil language, neither the fundamental linguistic resources are available in machine-readable form nor are the commercial prospects sufficiently large to warrant the considerable software development involved. Thus the prospects would seem bleak, except for a new research area of *statistical machine translation,* which offers hope that low-cost software which learns by example will enable at least reasonable translations to be obtained which would enable non-Tamil readers to understand the gist of documents written in Tamil.

Introduction

A 1997 random sample of web documents revealed that twelve percent were in a language other than English [Oard and Diekma 99]. Statistics in December 1999 from Excite show that, after the United States, the largest number of web pages (2 1/ million) originate from Japan. The Alta-Vista inclusion of a page translation option and a targeted language search capability shows a growing interest in non-English information sources. Cross language retrieval of European languages has now become relatively advanced (Ballesteros and Croft 1997). However, on the basis of population alone, Asian languages will provide a growing percentage of available information sources.

Commercial machine translation is also coming of age. One can now purchase low-cost software which will translate between English and the core European languages (French, German, Italian, Portuguese and Spanish), as well as Chinese and Japanese. Translation for web documents in many of these languages is available as an adjunct to search engines at Google and AltaVista (http://babelfish.altavista.com). Other translation sites are available for Russian (http://www.translate.ru) and recently Arabic (http://tarjim.ajeeb.com/ajeeb).

However, in this rush to machine translation, the languages of the Indian subcontinent have been forgotten. Seven of these languages (Hindi, Bengali, Marathi, Punjabi, Telugu, Tamil and Urdu) are among the twenty-five top languages in the world when ranked by number of people speaking that language. The commercial market for machine translation from these languages to English seems quite limited, since most newspapers in India have English editions in which all important stories also appear. And if the market for translation to English is not large enough to support the development of commercial machine translation, there seems little prospect for machine translation between Tamil and, say, Hindi.

I, as a non-Tamil reader with a more than passing interest in the Tamil Nadu region of India, would like to be able to read articles such as found in Figure 1: *News Article about Pakistani Nuclear Scientist*. How can I do this if there is no automatic machine translation capability available? The answer is that I cannot. However, I, as a researcher into search capabilities across languages know that research is underway to rapidly develop new translation software using the techniques of statistical machine translation.

Statistical Machine Translation

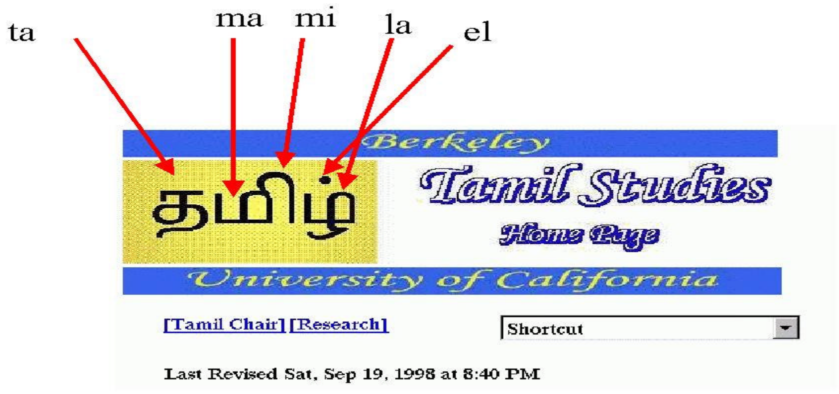
Statistical Machine Translation has its roots in a famous memorandum by Warren Weaver (1949) in which he observed:

> It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?

This memorandum led to early research and development in machine translation. However after early primitive successes it was recognized that machine translation could not be resolved by decoding methods alone and that fundamental research into the structure of natural language was needed. However, another limitation of that era was the weakness of computers in terms of raw processing power to deal with large amounts of text. In the next 40 years much progress was made in natural language understanding and in the division between syntax structure and semantic analysis. By the early 1990s, however, computational power had increased such that another look at the problem from a statistical decoding point of view could be attempted and was by P Brown and colleagues at IBM research (Brown et al 1990 , Brown et al 1993) who developed layered mathematical models of the translation process which could be trained statistically from what is known as parallel corpora. Parallel corpora are document collections which have been translated (by human beings) sentence-by-sentence from one language to another. The first such corpus to be widely used (and used by Brown et al) was the Canadian Hansards, records of the Canadian parliament which are made available in both English and French. Since the publication of the Brown papers there has been an explosion of research and prototyping of statistical machine translation. Other resources which have been experimented with include the United Nations texts which are often translated into the five official U.N. languages: Arabic, Chinese, English, French and Russian. Literally millions of sentences of electronic text are available as raw material for statistical machine translation in these languages. Outside of these five languages, however, parallel corpora are difficult to come by, and the focus of much current research is in bootstrapping statistical machine translation in the absence of large resources (Al-Onaizan et al 2000, Koehn and Knight 2001). Such will be the case for the application of statistical machine translation to the Tamil language.

Assembling necessary linguistic resources for Tamil

Two major problems exist in connection with machine translation and cross-language retrieval of Tamil (and other Indian languages). First is the lack of machine-readable resources for either machine translation or cross-language dictionary lookup. Most Tamil research has been in the humanities (http://tamil.berkeley.edu) and has focussed on the rich classical literature of the Tamil Nadu state of south India, an unbroken tradition of 2,000 years of creative writing. Since much of this literature has a religious focus, existing dictionaries have a comprehensive catalog of religious figures and forms as well as poetic metaphors. To be useful for WWW retrieval, such dictionaries need to be pared of poetic and classical terminology and augmented with modern words as found in recent newspaper and magazine texts. In addition, as the insert shows:



Last Revised Sat, Sep 19, 1998 at 8:40 PM

Tamil is a phonetic language. By default each Tamil consonent is followed by an ‘a’ vowel sound. The vowel sound is modified by the addition of glyphs, such as the curlicue following the ‘m’ which changes it sound from ‘ma’ to ‘mi’. To suppress the vowel sound one places a dot over the consonant. This means that borrowed words from English or other languages will ‘sound’ similar in Tamil to their native language (Malten 1996). We propose to search for such words by making the assumption that many words are "borrowed" from English by transliteration into the phonetic Tamil representation . We will apply the technique of "phonetic back-transliteration" of Knight [Knight and Graehl 1998, Stalls and Knight 1998] which has been demonstrated on Japanese and Arabic. Figure 2 is a story from Madras India about the Clinton-Yeltsin summit meeting for which phonetic recognition of the principals of the story is shown.

A Tamil Corpus and Tamil retrieval

We have assembled a corpus of Tamil news stories from July 1, 1998 through Feb 1, 1999 from the Thinaboomi site (http://www.thinaboomi.com). This corpus contains over 3,000 news stories in the Tamil language, and provides a rich source for modern Tamil linguistic studies and retrieval. About ten percent of these stories (all international news stories) were translated into English through the efforts of researchers at the University of Cologne (Koeln), Germany Institute of Indology and Tamil Studies. This corpus has been used to develop an experimental statistical machine translation system from Tamil to English by the Information Sciences Institute (http://www.isi.edu), one of the leading machine translation research organizations (German 2001). One of the tasks of the our future research will be to align this corpus and develop a general purpose bilingual dictionary using techniques already applied to Japanese-English scientific words (Chen, Gey et al 1999, Dunning 1993).

This corpus is but a small sample of what is needed to develop appropriate machine translation capabilities using statistical machine translation techniques. It is incumbent upon the Indian government to make available the electronic texts of all materials which have been translated (by humans) into more than one official Indian language. This can form the basis for training of a statistical translation engine to produce customized translation software which will translate between pairs of languages.

Summary

To our knowledge there is no direct machine translation software development being done on Indian subcontinent languages – languages spoken by nearly twenty percent of the world's people. The reasons are simple, the resources available for traditional machine translation are simply non-existent and the commercial viability of such development seems dim. However, the new field of research into statistical machine translation offers hope that Tamil translation may be just around the corner, to the pleasure and social benefit of all.

References

Al-Onaizan, Yaser, U. Germann, Ulf Hermjakob, K Knight, P Koehn, D Marcu and K Yamada, "Translating with Scarce Resources," In *Proceedings of the National Conference on Artificial Intelligence (AAAI),* 2000.

Brown, P, J Cocke, S Pietra, V D Pietra, F Jelineck, J D Lafferty, R Mercer and P S Roossin, 1990, "A Statistical Machine Approach to Machine Translation," *Computational Linguistics,* Vol 16, # 2, pp 79-85.

Brown, P, S Pietra, V D Pietra, and R Mercer, 1993, "The Mathematics of Statistical Machine Translation," *Computational Linguistics,* Vol 19, # 2, pp. 263-312.

Dunning, Ted 1993 "Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics*, vol 19, no. 2, pp.61-74.

A Chen, F Gey, K Kishida, H Jiang and Q Liang, "Automatic Construction of a Japanese-English Lexicon and its Application in Cross-Language Information Retrieval," presented and published in the *Proceedings of the Joint ACM DL'99/SIGIR'99 Workshop on Multi-Lingual Information Discovery and Access*, August 14, 1999, available at http://www.sims.berkeley.edu/research/metadata/papers/01chen99.ps

Germann, Ulrich, 2001 "Building a Statistical Machine Translation System from Scratch: How much Bang for the Buck Can We Expect," *Proceedings of the ACL 2001 Workshop on Data Driven Machine Translation,* Toulouse, France*,* July 7, 2001, 8pp.

Kando, N, and A Aizawa, "Cross-lingual Information Retrieval using Automatically Generated Multilingual Keyword Clusters," *Proceedings of IRAL'98, the Third International Workshop on Information Retrieval with Asian Languages, Singapore*, Oct 16-17, 1998.

Knight, K and J. Graehl "Machine Transliteration," Journal version *Computational Linguistics*, 24(4), 1998 (available at www.isi.edu/natural-language/mt/transliterate.ps)

Koehn, Philipp and Kevin Knight, "Learning a Translation Lexicon from Monolingual Corpora"

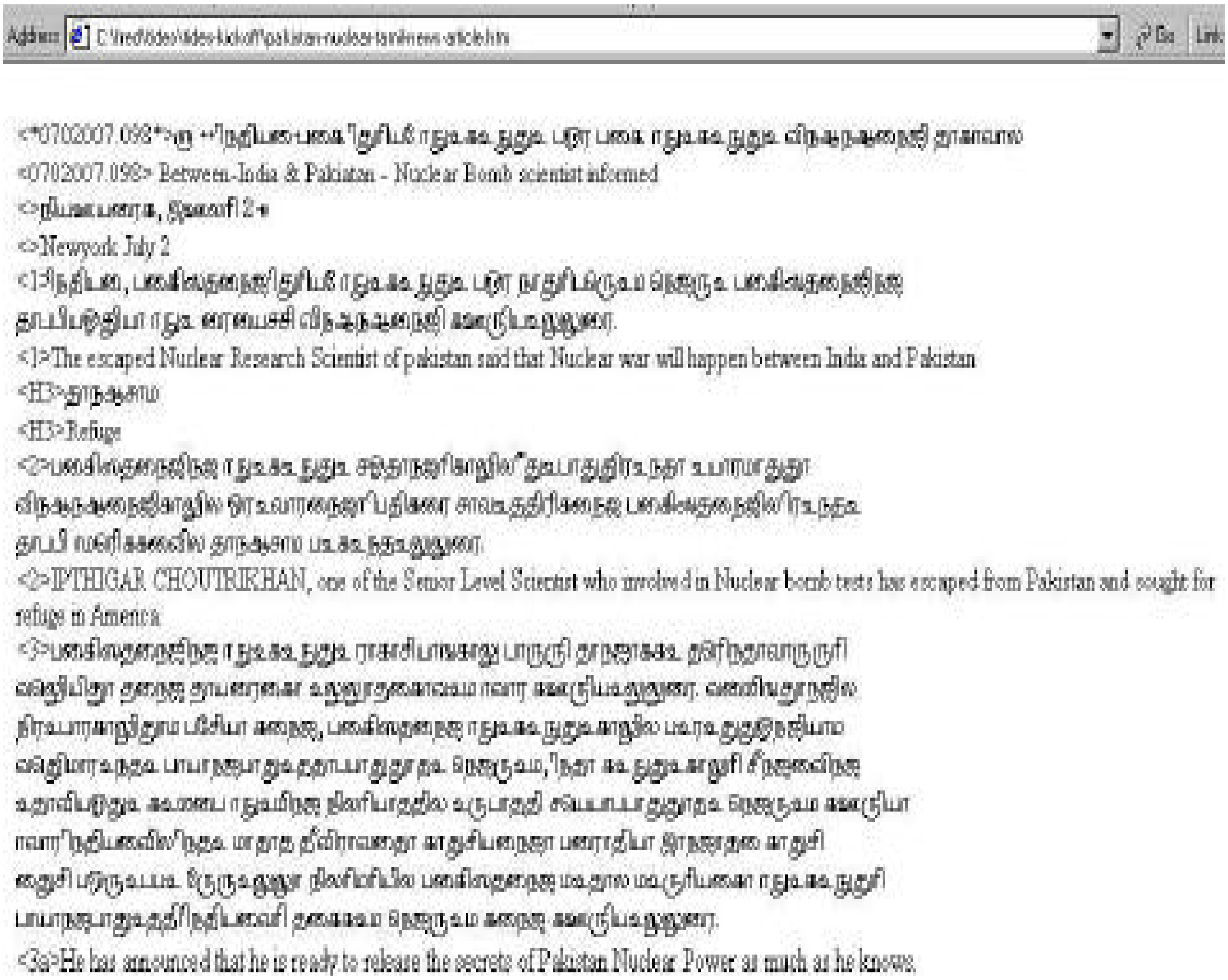
Oard D and A Diekema, ARIST review paper "Cross-Language Information Retrieval" To appear, **Annual Review of Information Science and Technology**, M Williams Editor, ASIS Press, 1999.

Stalls, B and K. Knight, "Translating Names and Technical Terms in Arabic Text,*", Proc of the*
COLING/ACL Workshop on Computational Approaches to Semitic Languages, 1998, available at www.isi.edu/natural-language/mt/arabic-translit.ps .

Weaver, Warren (1949) "Translation" in William N. Locke & A. Donald Booth (eds) (1955) *Machine Translation of Languages: Fourteen Essays*, The Technology Press of the Massachussetts Institute of Technology/John Wiley (New York)/Clapham & Hall (London), 1955. pages 15-23.

Yamada, Kenji and Kevin Knight, "A Syntax-based Statistical Translation Model",

Figure 1: News article about Pakistani Nuclear Scientist


Address
C:\fred\docs\ides-kickoff\pakistan-nuclear-tamil\news-article.htm
Go
Links

<*0702007.098*> [illegible]
<0702007.098> Between-India & Pakistan - Nuclear Bomb scientist infomed
<> [illegible] 2+
<>Newyork July 2
<1> [illegible]
[illegible]
<1>The escaped Nuclear Research Scientist of pakistan said that Nuclear war will happen between India and Pakistan
<H3> [illegible]
<H3>Refuge
<2> [illegible]
[illegible]
[illegible]
<2>IPTHIGAR CHOUTRIKHAN, one of the Senior Level Scientist who involved in Nuclear bomb tests has escaped from Pakistan and sought for refuge in America
<3> [illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
[illegible]
<3a>He has announced that he is ready to release the secrets of Pakistan Nuclear Power as much as he knows.

Figure 2: Phonetic recognition of proper nouns in Tamil