International Journal of Computer and Communication Technology

Volume 8 | Issue 2 Article 12

April 2017

COMPREHENSIVE APPROACH FOR BILINGUAL MACHINE TRANSLATION

SHARANBASAPPA HONNASHETTY
Doddappa Appa Institute of MCA, Gulbarga, India, saranbasappahonnasetty@gmail.com

DR. M. HANUMANTHAPPA
hanu6572@hotmail.com

COMPREHENSIVE APPROACH FOR BILINGUAL MACHINE TRANSLATION

SHARANBASAPPA HONNASHETTY[1], DR. M HANUMANTHAPPA[2]

[1]Doddappa Appa Institute of MCA, Gulbarga, India
[2]Bangalore University, Bangalore, India

Abstract- Machine Translation has been a major focus of the NLP group since 1999, the principal focus of the Natural Language Processing group is to build a machine translation system that automatically learns translation mappings from bilingual corpora. This paper explores a novel approach for phrase based machine translation from English to Kannada and Kannada to English. The source text is analyzed then simple sentences are translated using the rules and the complex sentences are split into simple sentences later translation is performed.

Keywords- *Natural Language Processing (NLP), Phrase Structure Grammar (PSG), Shift-Reduce Parsing (Tagging).*

I. INTRODUCTION

Perception and communication are the essential components of intelligent behavior; they provide the ability to interact effectively. Humans who reside at different regions have same perception but lack of communication language. To achieve effective interaction a system is required that translates sentences from one native language to another.

The syntactic structures related to finiteness are vastly different in the Dravidian languages, but these differences are not always visible on the surface of sentences and the complexity of differences and similarities between English and Kannada might yield additional insights regarding the status of cross-linguistic influences in the development of finiteness under simultaneously bilingual conditions.

The aim of this paper is to investigate the generalized translation from English to Kannada and Kannada to English sentences and exploring the motivations for uncommon variations, this paper give a short description of the structural features of English and Kannada languages.

II. REVIEW OF LITERATURE

A. shift-reduce parsering

A shift-reduce parser starts out with the entire input string and looks for a substring that matches the right-hand side of a production. If one is found, the substring is replaced by the left-hand side symbol of the production. Reductions occur until the string is reduced to just the start symbol.

A shift reduce parsing begins with actual words appearing in the sentence therefore it is a data driven, in a shift-reduce parser it tries to find sequences of words and phrases that correspond to the right hand side of a grammar production, and replace them with the left-hand side, until the whole sentence is reduced to an S.

B. Phrase structure of English and Kannada Language

English is essential language that maintains the SVO word order through all levels of syntactic complexity. In the absence of auxiliaries or modal verbs, subject-verb agreement or past tense affixes attach to the main verb, but the verb does not move away from its base position at the beginning of the Verb Phrase. In contrast to it the Kannada language is highly agglutinative language with three gender forms namely masculine, feminine and neutral, with this Word order plays an important role in positional languages like English which normally follow right-branching with Subject-Verb-Object orders where as In Kannada language is verb final language and all the noun phrases in the sentence normally appear to the left of the verb, hence it is 'Left branching language' and the adjectives, genitive and relative clauses precede their head nouns in a sentence. The subject noun phrase may also appear in many different positions relative to other noun phrases in the sentence.

E.g. in a sentence "Rama went to school" depict the structure of English language (i.e. Subject-verb-object) and its equivalent Kannada sentence is "rAma SAlEge hOdanu" (gÁªÀÄ ±Á¯ÉÃUÉ ºÉÆÃzÀ£ÀÄ) which depicts the structure of Kannada language (i.e. Subject-Object-Verb).

C. Verb morphology in south Dravidian Languages

The PNG and the tense marker concatenated to the verb stems are the two aspect of verb morphology in South Dravidian languages. The verbal inflectional morphemes attach to the verbs providing information about the syntactic aspects like number, person, case-ending relation and tense. The PNG features of the head noun of the subject NP determine the agreement marker of the verb. English language has Inflectional and Derivational Morphemes and the English has only eight inflectional affixes:

{PLU} = plural Noun -s boys

{POSS} = possessive Noun -'s boy's
{COMP} = comparative Adj -er older
{SUP} = superlative Adj -est oldest
{PRES} = present Verb -s walks
{PAST} past Verb -ed walked
{PAST PART} = past participle Verb -en driven
{PRES PART} = present participle Verb -ing driving
Notice that, as noted above, even irregular forms can be represented morphologically using these morphemes. E.g. the irregular plural sheep is written as {sheep} + {PLU} and there is an indefinite number of derivational morphemes, the following are some derivational suffixes:
{ize} attaches to a noun and turns it into a verb: rubberize
{ize} also attaches to an adjective and turns it into a verb: normalize
{ful} attaches to a noun and turns it into an adjective: playful, helpful
{ly} attaches to an adjective and turns it into an adverb: grandly, proudly etc.

The Kannada language has various PNG suffixes that can be attached to any Kannada verb root word and the table 2.1 shows [3] these suffixes

TABLE I. PNG- SUFFIXES IN KANNADA

P	N	G	PNG Suffix			
			Present	Future	Past	Contingent
1st	S	M/F	K£É (Ene)	J£ÀÄ, J (enu,e)	J£ÀÄ, J (enu,e)	J£ÀÄ (Enu)
	P	M/F	KªÉ (Eve)	KªÀÅ (Evu)	KªÀÅ (Evu)	KªÀÅ (Evu)
2nd	S	M/F	F,F0iÉÄ (I,Iye)	F,E0iÉÄ (I,iye)	E0iÀÄ (iya)	F0iÀÄ (Iya)
	P	M/F	Fj (Iri)	Fj (Iri)	Ej (iri)	Fj (Iri)
3rd	S	M	D£É (Ane)	C£ÀÄ (anu)	C£ÀÄ (anu)	C£ÀÄ (anu)
	S	F	D¼É (ALe)	C¼ÀÄ (aLu)	C¼ÀÄ (aLu)	C¼ÀÄ (aLu)
	P	M/F	DgÉ (Are)	DgÀÄ (Aru)	CgÀÄ (aru)	DgÀÄ (Aru)
	S	N	EzÉ (ide)	GzÀÄ (udu)	EvÀÄ (itu)	FvÀÄÛ (Ittu)
	P	N	EªÉ (ive)	DªÀÅ (Avu)	CªÀÅ (avu)	DªÀÅ (Avu)

P: Person N:Number G:Gender
S: Singular P:Plural M:Masculine F:Feminine N:Neuter

All the verb words use the same present and future tense markers but all the South Dravidian languages uses different past tense markers based on the types of verb paradigms. The table 2 shows the different tense markers that are used in Kannada language [3]

TABLE II. TENSE MARKERS IN KANNADA

Tense	Tense Markers
Present	utt(GvïÛ)
Past	tt(vïÛ),MMt(AAvï),t(vï),d(zï), dd(zïÝ),id(Ezï),MMd(AAzï),D(qï),T(mï), k(PïÌ),MMD(AAqï)
Future	uv(Gªï)

III. METHODOLOGY

Sentences are composed of groups of words making up phrases. Phrases may contain other phrases. Phrases fall into a small set of types, the most important of which are NP (noun phrase), PP (prepositional phrase) and VP (verb phrase). Every phrase has a 'head' word which defines its type. A simple English sentence S is composed of a noun phrase Followed by a verb phrase. In this paper a novel approach is given to translate English sentences which comprise PP to equivalent Kannada sentences and vice versa approach can also be performed by using the same steps.

The below architecture show the flow of operations performed during the translation of an English to Kannada sentences and reverse process can be used to translate from Kannada to English sentences.

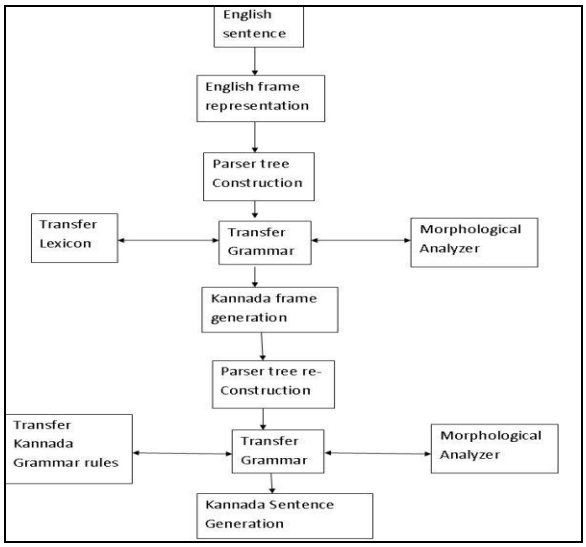

Figure 3.1 Architecture of translation system

The following steps are designed to produce a novel approach in translating sentences from English to Kannada and Kannada to English by using rule-based method. The outline of the our methodology is as follows

- Parsing and structural representation.
- Morphological analysis.
- Lexicon and mapping.
- Translation.

A. Parsing and Structural Representation

In this method we use shift reduce parsing technique to construct a parse tree [9] where a shift-reduce parser tries to find sequences of words and phrases that correspond to the left hand side of a grammar production, and replace them with the right-hand

side, until the whole sentence is reduced to an S and the compound sentences are split with respect to conjunction by generating individual simple sentences later build constituent structures for all individual sentences.

The syntax of a language can be described by a 'formal grammar' which consists of

- A set of non-terminal symbols
- A start symbol
- A set of terminal symbols (words)
- A set of productions (also called re-write rules)

Nonterminal symbols:

S = sentence	PP = prepositional phrase
NP = noun phrase	DET = determiner
SNP = simple noun phrase	ADJ = adjective noun
VP = verb phrase	PV = preposition verb
V = Verb	PRN = pronoun
VC = verb complements	N=Noun

The following production rules used by the parser to make the parser tree structure for English sentences

S -> NP VP
NP -> DET SNP | PRN
SNP -> ADJ N
VP -> V VC
VC-> NP PP
PP -> PV NP

Finally the transfer rule was used to change the structure of English sentence according to Kannada chunk order as shown below

S -> NP VP
NP -> DET SNP| PRN
SNP -> ADJ N
VP -> VC V
VC-> NP PP
PP -> NP PV

After applying the transfer rules, the reordered parse tree look as shown in figure 3.2

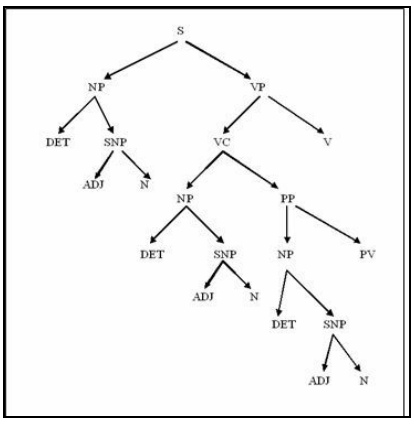

Figure 3.2 Tree Structure after reordering

B. Morphological analysis

In this method a porter stemming algorithm can be used to reduce the English words to morphemes, the inflectional affixes and derivational suffixes are also reduced by porter's algorithm

C. Lexicon and mapping

In our approach lexicon is arranged in alphabetical order and Keep a separate lexicon that contains frequently used words as well as a domain specific components of words can be kept in different lexicon to increase the efficiency of search process [6]. Once the word is matched in the lexicon, it is tagged to respective Kannada equivalent word which is defined in the lexicon which is in the same language using one to one mapping technique and the following algorithm [1] can be used to perform tagging

Input: Untagged English Sentence
Output: Tagged Translated Kannada Sentence
Tag<= First Word in Sentence
For each word in the Sentence Do
If the Word is tagged
Stop
Else
Tag<=Word
End if
End For
Return Tagged Translated Kannada Sentence

D. Translation

There are a number of distinct approaches to automating translation between NLs. The most obvious approach is to use a SINGLE intermediate representation, e.g. a syntax tree with extra information added. For example, translation from English to Kannada could start with an English grammar, extended to generate syntax trees. Inputting an English sentence using an English lexicon produces a syntax tree representing the English sentence. The syntax tree, as noted above, should store the 'base' word (i.e. lexeme) plus key properties. The base words in the tree can then be mapped to their Kannada equivalents using an English/Kannada translation lexicon, with key properties. The resulting syntax tree for Kannada is then put back through a Kannada grammar and a Kannada lexicon to yield a Kannada sentence

Once mapping (tagging) is completed, the resultant Kannada sentence is obtained which is not in proper meaning so next step is to apply some rules to make it a meaningful Kannada sentence. If the input English sentence is a simple sentence then these rules can be ignored. A few rules are listed here that are applied when an English sentence consists of Prepositional Phrase and these are known to be inflections to the noun stem.

Rule 1: add suffix "annu" (C£ÀÄß) to noun phrase of the VP root.

Rule 2: add suffix “da” (CzÀ) to noun phrase of the PP root when “with” preposition encountered.
Rule 3: add suffix “ige” (EUÉ) to noun phrase of the PP root when “to” preposition encountered.
Rule 4: add suffix “alli” (C°è) to noun phrase of the PP root when “in” preposition encountered.

E.g. A simple translation of an English Sentence “The dog saw a man in the park” to equivalent Kannada sentence “nAyi oMdu manuShyanannu pArkinalli nODitu” (£Á¬Ä MAzÀÄ ªÀÄ£ÀÄµÀå£À£ÀÄß ¥ÁQð£À°è £ÉÆÃrvÀÄ)

IV. CONCLUSION AND FUTURE WORK

In this paper we have discussed how Machine translation from English to Kannada can be achieved with the use of prepositional phrase (PP). Part of Speech Taggers are used for resolving the ambiguity in the Sentence for translation for English to Kannada and it also help for resolving the problem of Sentence format such as English have Subject-Verb-Object Format where as Kannada have Subject-Object- Verb format. In future this will help to develop the other modules for Different Indian Languages as well as for other languages

REFERENCES

[1] POS Tagger for Kannada Sentence Translation , Mallamma V Reddy, Dr.M. Hanumanthappa, International Journal of Emerging Trends & Technology in Computer Science(IJETTCS) Volume 1, Issue 1, May-June 2012

[2] Natural Language Processing K.R. Chowdhary Professor & Head CSE Dept. M.B.M. Engineering College, Jodhpur, India April 29, 2012

[3] “A Novel Approach for English to South Dravidian Language Statistical Machine Translation System”, Unnikrishnan P, Antony P J, Dr Soman K P, International Journal on computer Science and engineering (IJCSE) Vol 02, No. 08, 2010,2749-2759.

[4] K. Narayana Murthy, “Computer Processing of Kannada Language”, University of Hyderabad.

[5] R.M.K. Sinha and Anil Thakur, “Synthesizing Verb Form in English to Hindi Translation”, Case of Mapping Infinitive and Gerund in English to Hindi, Proceedings of International Symposium on Machine Translation, NLP and Translation Support System (iSTRANS- 2004), November 17-19, 2004, Tata Mc Graw Hill, New Delhi, pp: 52-55.

[6] Amitabha Mukerjee, Achla Raina, Pankaj Goyal,and Pushpraj Shukla, A unified Computational Lexicon for Hindi-English code-switching Proceedings International Conference on Natural Language Processing (ICON), Hyderabad, India, December 19-22, 2004.

[7] Blaheta, Don, and Eugene Charniak. 2000. Assigning function tags to parsed text. In Proceedings of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics, pages 234–240, Seattle.

[8] Hai Zhao, Yan Song, Chunyu Kit, and Guodong Zhou. 2009. Cross language dependency parsing using a bilingual lexicon. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1 (ACL '09), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 55-63.

[9] B.S. Baker 1979. Composition of top-down and bottom-up tree transductions Inform. and Control, 41(2):186–213

[10] K. Knight. 2007. Capturing practical natural language transformations. Machine Translation 21, 121–133.

[11] Takuya Matsuzaki, Yusuke Miyao and Junichi Tsujii. 2007. Efficient HPSG Parsing with Supertagging and CFG-filtering. Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007).

[12] Johanne Paradis, “Bilingual Children’s Acquisition of English Verb Morphology: Effects of Language Exposure, Structure Complexity, and Task Type”. A journal of research in Language studies, ISSN 0023-8333.

[13] Dr Susanne Döpke, “The Development of Verb Morphology and the Placement of Finite Verbs in Young Bilingual German-English-Speaking Children”.

[14] Eric Potsdam,” English Verbal Morphology and VP Ellipsis” , the Proceedings of the Twenty-Seventh Annual Meeting of the North East Linguistic Society. Amherst, Mass.: GLSA, University of Massachusetts at Amherst, 353-368.

[15] Dr.S.Saraswathi,P.Kanivadhana,M.Anusiya,S.Sathiya,”BILI NGUAL TRANSLATION SYSTEM” International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397 Vol. 3 No. 3 Mar 2011.

◈◈◈