Development of Punjabi Textgenerator for Translation from English

Introduction:

AnglaBharati system architecture exploits structural similarity of Indian languages. This structural similarity is more homogeneous within each family of languages such as within the Indo-Aryan family, Dravidian family and others. AnglaBharati system translates the English text into an intermediate language that follows the structure of the target language family of Indian languages. This intermediate representation contains most of the semantic data needed to construct the final target language text. This intermediate language is PLIL (Pseudo Lingua for Indian Languages) which is generated by rulebase of the system, also which is an input to text generator that converts the pseudo code in PLIL into the target language. The main task in the module is to generate TL words from the morphological information available in the PLIL. It also constructs TL sentences, such that the syntactic and semantic constraints of target language are met. Morphological and syntactic rules, which covers all the features of target language, are required as input to this module. This paper shows the design architecture of module developed in AnglaPunjabi that decodes the encodings available in PLIL to generate Punjabi translations. This module is named text generator.

Components of PLIL:

PLIL is similar to Inter-lingua of an MT System using Inter-lingua approach, the difference is that it only caters to a class of languages for which it has been designed for and in AnglaBharati English is the source language. The English to PLIL encoder generates a structure as per Punjabi language and text generator acts a s decoder to generate target language structure. For PLIL generation, engine contains rules for mapping structures of sentence from English to Indian languages. This is the translation engine, which identifies the structure of English sentences and transforms that into PLIL. Different patterns of the source language have been captured by examining the corpus of the source language. Since many of the Indian languages are similar in nature, an intermediate code or a pseudo-target catering to a class of target languages is generated.

PLIL consists of 2 major components:

i. A bilingual lexical database of English to Punjabi: An English rootword/lexicon is mapped on to Punjabi lexicon along with its associated grammatical and semantic information. A root word may have multiple parts of speech or multiple meanings.

ii. A grammar representing Indian language structure, specifically Punjabi in this case. As AnglaPunjabi is adaptation of AnglaBharati methodology, so the grammar defined also goes well for Punjabi language. This grammar has been loosely defined around a CFG formalism which generates the word order for the class of languages. A sentence in PLIL is defined in terms of NP, VP and other constructs as expected in

case of any natural language.

Processes involved in Text Generator:

(1) Paradigm file generation

With the help of paradigm files, root word is extracted from the original word and all the information about that word is retrieved. The paradigm number has to be generated for different word categories as per the requirement. The relationship between the adjacent words and hence the total sentence meaning is identified by the different syntactic information. This syntactic information is identified from the paradigm number assigned to certain word category.

1.1 Noun paradigm: The Punjabi nouns are classified based on gender, number, case markers and word endings. The gender (masculine, feminine), number (singular,plural) and case markers(direct,oblique) are combined together to produce forms along with a paradigm number. The paradigm number is a unique id given to a set of those nouns that have similar word ending and generate forms by substituting suffix.

Example 1: Root word- boy (masculine)

Direct	Oblique
ਮੁੰਡਾ (muMdA)	ਮੁੰਡੇ (muMde)
ਮੁੰਡੇ (muMde)	ਮੁੰਡਿਆ (muMdiAM)

Root	Number	Case	Generated Form
muMd_A	singular	direct	muMd_A
muMd_A	plural	direct	muMd_e
muMd_A	singular	oblique	muMd_e
muMd_A	plural	oblique	muMd_iAz

For another noun ending with "_A", will have same paradigm if the generated forms for two cases are having similar variations (_A,_e,_e,_iAz). For example, GodZA (masculine) is "_A" ending and its variations are GodZA, GodZe, GodZe, GodZiAz similar to muMdA. So both words will share a common paradigm. In Angla Punjabi, 10 feminine and 17 masculine noun paradigms exists.

1.2 Verb Paradigm : For Punjabi verbs the forms are inflected according to gender, number, person and tenses. Transitive and Intransitive nature of verbs are also considered. The following example shows the various formations of a paradigm.

mple 3: Root verb KA / खा

Tense	G	N	Person	PunjabiVerb Forms(Gur/WX)
Past tense	m	s	3rd	ਖਾਧਾ (KAxA)
	m	p	3rd	ਖਾਧੇ (KAxe)
	f	s	3rd	ਖਾਧੀ (KAxI)
	f	p	3rd	ਖਾਧੀਆਂ (KAxIAz)
Subjunctive	m	s	1st	ਖਾਵਾਂ (KAvAz)
	m	s	2nd	ਖਾਵੇ (KAvez)
	m	s	3rd	ਖਾਵੇ (KAve)
	m	p	1st	ਖਾਈਏ (KAIye)
	m	p	2nd	ਖਾਵੇ (KAvo)
	m	p	3rd	ਖਾਣ (KANa)
future	m	s	1st	ਖਾਵਾਂਗਾ (KAvAzgA)
	m	s	2nd	ਖਾਵੇਗਾ (KAvezgA)
	m	s	3rd	ਖਾਵੇਗਾ (KavegA)
	m	p	1st	ਖਾਵਾਂਗੇ (KAvAzge)
	m	p	2nd	ਖਾਓਗੇ/ਖਾਵੇਗੇ (KAoge/KAvoge)
	m	p	3rd	ਖਾਣਗੇ (KANage)

	f	s	1st	ਖਾਵਾਂਗੀ (KAvAzgI)
	f	s	2nd	ਖਾਵੇਂਗੀ (KAvezgI)
	f	s	3rd	ਖਾਵੇਗੀ / ਖਾਏਗੀ (KAvegI)
	f	p	1st	ਖਾਵਾਂਗੀਆਂ (KAvAzgIAz)
	f	p	2nd	ਖਾਵੇਂਗੀਆਂ (KAvogIAz)
	f	p	3rd	ਖਾਣਗੀਆਂ (KANagIAz)

mple 3: Root ver

Similar to noun, verbs similar in nature are recognized as common paradigm. AnglaPunjabi system has 14 verb paradigms.

1.3 Adjective paradigm: Adjectives in Punjabi has variations only for words ending with "_A", so a single paradigm has been created for two genders. The forms are generated by considering direct/oblique forms along with gender and number information.

Eg.
Root Word: good

Punjabi-masculine		Punjabi- feminine	
Direct	Oblique	Direct	Oblique
ਚੰਗਾ (ਚੰਗਾ)	ਚੰਗੇ (ਚੰਗੇ)	ਚੰਗੀ (ਚੰਗੀ)	--
ਚੰਗੇ (ਚੰਗੇ)	ਚੰਗਿਆਂ (ਚੰਗਿਆਂ)	ਚੰਗੀਆਂ (ਚੰਗੀਆਂ)	ਚੰਗੀਆਂ (ਚੰਗੀਆਂ)

Root	Gender	Number	Case	Generated Form
caMgA	"m"	"s"	"d"	caMg_A
caMgA	"m"	"p"	"d"	caMg_e
caMgA	"m"	"s"	"o"	caMg_e
caMgA	"m"	"p"	"o"	caMg_iAM
caMgA	"f"	"s"	"d"	caMg_I
caMgA	"f"	"p"	"d"	caMg_IAM
caMgA	"f"	"s"	"o"	caMg_I
caMgA	"f"	"p"	"o"	caMg_IAM

1. 4 Paradigm for pronouns: A pronoun is a kind of noun, but its function is different from noun. One of its classifications is described below:

Root	del char	Generated Form
wusIz	3	wusIz
wusIz	3	wuhAnUM
wusIz	3	wuhAde woz
wusIz	3	wuhAde viYca
wusIz	3	wuhAde ‘we
wusIz	3	wuhAde
wusIz	3	wuhAdI
wusIz	3	wuhAde waYka

(2) Postposition disambiguator

Punjabi postpositions are similar to prepositions in English. These link noun, pronoun, and phrases to other parts of the sentence. Some Punjabi postpositions are ਨੇ nē, ਨੂੰ nūṃ, ਉੱਤੇ uttē ‘over’, ਦਾ dā ‘of’, ਕੋਲੋਂ kōlōṃ ‘from’, ਨੇੜੇ nēḍaē ‘near’, ਲਾਗੇ lāgē ‘near’ etc. In Punjabi, postpositions follow the noun or pronoun unlike English, where these precede the noun or pronoun, and thus termed prepositions. This module maps monosemous or polysemous prepositions in English with lexical postpositions in Punjabi.

Example1: A girl *with* beautiful eyes. < > sohaNI | KUbasUrawa ~ aVKa *vAlI* ^ ika | {} ~ ladZakI

Example2: The kid is playing *with* letters. <> baccA aVKara *nAla* Keda rihA hE

Here, English preposition “with” has been disambiguated with two meanings in Punjabi, “vAlI” and “nAla”. Hence, suitable addendum has been substituted in Punjabi for “with” by comparing the English sentence with its Punjabi translation.

(3) Generation of Verb forms

This module derives the verb forms in Punjabi using TAM (Tense, Aspect and Modality). This module is used to find out proper translation of the sentence with the proper suffix. The design has five fields:

a) Finiteness

b) Auxiliary Verb

c) Main verb type

d) Phrasal field

e) Suffix

Example1: “normal”, “am”, verb_5,-1, “_ rihA_*hAz

In this “rihA_*hAz” is suffix. * before “hAz” means this word cannot change in Punjabi after translation.

English: I am playing.

Punjabi: mEz Keda rihA hAz.

Similarly, consider the following sentences:

Example 2(a): He went to the market < > uha bajZAra giA

Example 2(b): He is going to the market. < > uha

bajZAra jA rihA hE

Example 2(c): They want to go. < > uha jANa xe laI iYCAxe hana

Example 2(d): Can I go ?< > kI mEz jA sakaxA hAz

Example 2(e): I got my jacket cleaned. < > mEz ne ApaNI jEkata saPZa karavAxA

Example 2(f): She made her children do their homework. < > uha usaxiAM baVciAz nUm unhAz Gara xA kaMma karaNa xe laI majabUra kIwI

We notice that in (2 a) and (2 b), the difference in the formation of the verb (go) irrespective of the same structure of sentence. Here the type of the verb has been used for deciding the correct form. However, in (2 c) the type of the verb and the pattern type are used to determine the desired formation. In (20 d) the type of the verb, pattern type and the auxiliary has to be considered for its correct formation. However in (2 e) and (2 f) causative verb (get, make) are taken into consideration for appropriate formation.

(4) Disambiguation of "to-infinitive"

The semantic distribution of a single preposition will be varying in different context due to the influence of nouns and main verbs that follow. For instance, a preposition *to* can have multiple mapping patterns in Punjabi

1) [to = nUM]

The procession goes [*to* Kottayam] . < > jalUSa kotavama *nUM* jAxA hE.

NP,place

2) [to = nAla]

I have spoken [*to* him] already. < > mEz pahilAz hI usanUM *nAla* bola cuVkA hAz . NP, human

3) [to = xe vaVla]

I am going *to* the king.

mEz rAjA *xe vaYla* jA rihA hAz

4) [to = nUM/{}]

Please listen to him.

kripA usanUM nUM suNo

5) [to = waka]

He is going *to* the meeting.

uha mItiMga *waYka* jA rihA hE

The above examples show that the preposition meanings are derived by the semantic type of the main verb and that of the nominal elements that occur with the prepositions. Also it could be seen that the multiple postposition patterns are derived using the semantic category of verb and noun.

(5) Resolution of gerunds / participles

Gerund is a word that functions as a noun. It is derived by adding '–ing' to the end of a verb (jog), e.g. "Jogging is a good way of exercising".

Example 1: gerund as a verb

The gate needs repainting. < > xaravAjZe ko xubArA raMganA lodZa hE

In this example, the gerund is repainting which is a verb form acting as noun and in Punjabi this verb type has been used to get the correct mapping pattern .

Example 2: I like to sing a song < > mEz ika gANe nUM gANA pasaMxa karaxA hAz.

Here, to sing is infinitive clause and replaced by "nUM gANA" in Punjabi.

(6) Lexical choice for adjectives and adverbs

This module disambiguates the meanings of adjectives and adverbs.

(7) Symbol mapping

This module does mapping of English symbols/ keywords to Punjabi keywords. It also assigns a paradigm number, which tells how the Punjabi meaning changes its form. Basically this module maps different notations used in intermediate form. Following convention for paradigm numbers have been used:

English Keyword	Punjabi keyword	Paradigm No.
# "had_been"	rihA_sI	1

Consider a sentence, "Being paralytic, he did not go."Its Punjabi translation is "aXarazga hoNa xe kAraNa , uha nahIz giA". Here "hoNa xe kAraNa" is the mapping pattern substituted for adjectice "being".. like:

"adjbeing"",hoNa xe kAraNa",7

Another example,

(1) "adj_yet": He is good **yet** unsuccessful. uha ^ caMgA | vaXIA ~ **Pera BI** asaPala hE

(2) "had_been_never": We **had been** never friends asIz **kaxe nahIz** xoswa rihe sana