Word Disambiguation in Shahmukhi to Gurmukhi Transliteration

Tejinder Singh Saini
ACTDPL, Punjabi University,
Patiala, Punjab-147 002, India
tej74i@gmail.com

Gurpreet Singh Lehal
DCS, Punjabi University,
Patiala, Punjab-147 002, India
gslehal@gmail.com

Abstract

To write Punjabi language, Punjabi speakers use two different scripts, Perso-Arabic (referred as Shahmukhi) and Gurmukhi. Shahmukhi is used by the people of Western Punjab in Pakistan, whereas Gurmukhi is used by most people of Eastern Punjab in India. The natural written text in Shahmukhi script has missing short vowels and other diacritical marks. Additionally, the presence of ambiguous character having multiple mappings in Gurmukhi script cause ambiguity at character as well as word level while transliterating Shahmukhi text into Gurmukhi script. In this paper we focus on the word level ambiguity problem. The ambiguous Shahmukhi word tokens have many interpretations in target Gurmukhi script. We have proposed two different algorithms for Shahmukhi word disambiguation. The first algorithm formulates this problem using a state sequence representation as a Hidden Markov Model (HMM). The second approach proposes n-gram model in which the joint occurrence of words within a small window of size ± 5 is used. After evaluation we found that both approaches have more than 92% word disambiguation accuracy.

1 Introduction

1.1 Shahmukhi Script

Shahmukhi is a derivation of the Perso-Arabic script used to record the Punjabi language in Pakistan. Shahmukhi script has thirty eight letters, including four long vowel signs Alif ا[ə], Vao و[v], Choti-ye ی[j] and Badi-ye ے[j]. Shahmukhi script in general has thirty seven simple consonants and eleven frequently used aspirated consonants. There are three nasal consonants (ںٛ[ɳ], ن[n], م[m]) and one additional nasalization sign, called Noon-ghunna ں [ɲ].

In addition to this, there are three short vowel signs called Zer ِ[ɪ], Pesh ُ[ʊ] and Zabar َ[ə] and some other diacritical marks or symbols like hamza ٔ[ɪ], Shad ّ, Khari-Zabar ٰ[ɘ], do-Zabar ً[ən] and do-Zer ٍ[ɪn] etc. Arabic orthography does not provide full vocalization of the text, and the reader is expected to infer short vowels from the context of the sentence. Any machine transliteration or text to speech synthesis system has to automatically guess and insert these missing symbols. This is a non-trivial problem and requires an in depth statistical analysis (Durrani and Hussain, 2010).

1.2 Gurmukhi Script

The Gurmukhi script, standardized by Guru Angad Dev in the 16th century, was designed to write the Punjabi language (Sekhon, 1996); (Singh, 1997). It was modeled on the *Landa* alphabet. The literal meaning of "Gurmukhi" is *from the mouth of the Guru*. The Gurmukhi script has syllabic alphabet in which all consonants have an inherent vowel. The Gurmukhi alphabet has forty one letters, comprising thirty eight consonants and three basic vowel sign bearers. The first three letters Ura ੳ[ʊ], Aira ਅ [ə] and Iri ੲ[ɪ] of Gurmukhi alphabet are unique because they form the basis for vowels and are not consonants. The six consonants are created by placing a *dot* at the foot (pair) of the existing consonant. There are five nasal consonants (ਙ[ɲə], ਞ[ɲə], ਣ[ɳ], ਨ[n], ਮ[m]) and two additional nasalization signs, bindi ਂ [ɲ] and tippi ਁ [ɲ] in Gurmukhi script. In addition to this, there are nine dependent vowel signs (or diacritics) (ੁ[ʊ], ੂ [u], ੋ[o], ਾ[ə], ਿ[ɪ], ੀ[i], ੇ[e], ੈ[æ], ੌ[Ↄ]) used to create ten independent vowels (ੁ

Proceedings of the 9th Workshop on Asian Language Resources, pages 79–87,
Chiang Mai, Thailand, November 12 and 13, 2011.

[ਪ], ਊ [u], ਓ [o], ਅ [ə], ਆ [ɑ], ਇ [ɪ], ਈ [i], ਏ [e], ਐ [æ], ਔ [ɔ]) with three bearer characters: Ura ਉ[ਪ], Aira ਅ [ə] and Iri ਏ[ɪ]. With the exception of Aira ਅ [ə] independent vowels are never used without additional vowel signs. The diacritics which can appear above, below, before or after the consonant they belong to, are used to change the inherent vowel and when they appear at the beginning of a syllable, vowels are written as independent vowels. Some Punjabi words require consonants to be written in a conjunct form in which the second consonant is written under the first as a subscript. There are three commonly used subjoined consonants as shown here Haha ਹ[h] (usage ਨ[n] + ◌੍ + ਹ[h] = ਨ੍ਹ [n^{h}]), Rara ਰ[r] (usage ਪ[p] + ◌੍ + ਰ[r] =ਪ੍ਰ [pr^{h}]) and Vava ਵ[v] (usage ਸ[s] + ◌੍ + ਵ[v] = ਸ੍ਵ [sv]).

1.3 Transliteration and Ambiguity

To understand the problem of word ambiguity in the transliterated text let us consider a Shahmukhi sentence having total 13 words out of them five are ambiguous. During transliteration phase our system generates all possible interpretations in target script. Therefore, with this input the transliterated text has supplied all the ambiguous words with maximum two interpretations in the Gurmukhi script as shown in Figure 1.

اس دور وچ سبھ توں طاقتور اتے چلاک ویکتی قبیلے دا مکھی رہا

ਇਸ ਦੌਰ ਵਿਚ ਸਭ ਤੋਂ ਤਾਕਤਵਰ ਅਤੇ ਚਲਾਕ ਵਿਅਕਤੀ ਕਬੀਲੇ ਦਾ ਮੁਖੀ ਰਿਹਾ

is daur vic sabh tōṃ tākatvar atē calāk viaktī kabīlē dā mukhī rihā

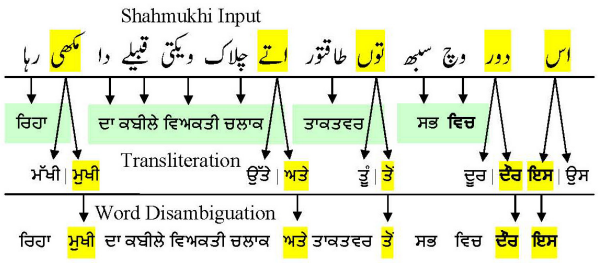

Figure 1. Word level Ambiguity in Transliterated Text

In a bigram statistical word disambiguation approach, the probability of co-occurrence of various alternatives such as <bos> ਇਸ |<bos> ਉਸ, ਇਸ ਦੌਰ | ਉਸ ਦੌਰ, ਇਸ ਦੂਰ | ਉਸ ਦੂਰ, ਦੌਰ ਵਿਚ | ਦੂਰ ਵਿਚ, ਸਭ ਤੋਂ | ਸਭ ਤੂੰ, ਤੋਂ ਤਾਕਤਵਰ | ਤੂੰ ਤਾਕਤਵਰ, ਤਾਕਤਵਰ ਅਤੇ | ਤਾਕਤਵਰ ਉੱਤੇ, ਅਤੇ ਚਲਾਕ | ਉੱਤੇ ਚਲਾਕ, ਦਾ ਮੁਖੀ | ਦਾ ਮੱਖੀ, and ਮੁਖੀ ਰਿਹਾ | ਮੱਖੀ ਰਿਹਾ are examined in the training corpus to estimate the likelihood. If the joint co-occurrence of following <bos> ਇਸ, ਇਸ ਦੌਰ, ਦੌਰ ਵਿਚ, ਸਭ ਤੋਂ, ਤੋਂ ਤਾਕਤਵਰ, ਤਾਕਤਵਰ ਅਤੇ, ਅਤੇ ਚਲਾਕ, ਦਾ ਮੁਖੀ, and ਮੁਖੀ ਰਿਹਾ bigram tokens are found to be more likely then the disambiguation will decide ਇਸ, ਦੌਰ, ਤੋਂ, ਅਤੇ and ਮੁਖੀ respectively as expected. Unfortunately, due to limited training data size or data sparseness it is quite probable that some of the alternative word interpretations are missing in the training corpus. In such cases additional information about word similarity like POS tagger and thesaurus may be helpful.

1.4 Causes of Ambiguity

The most common reasons for this ambiguity are missing short vowels and the presence of ambiguous character having multiple mappings in Gurmukhi script.

Sr	Word without diacritics	Possible Gurmukhi Transliteration
1	گل	ਗੱਲ /gall/, ਗਿੱਲ /gill/, ਗੁੱਲ /gull/, ਗੁਲ /gul/
2	تک	ਤਕ /tak/, ਤੱਕ /takk/, ਤੁਕ /tuk/
3	مکھی	ਮੱਖੀ/makkhī/, ਮੁਖੀ/mukhī/
4	ہن	ਹਨ /han/, ਹੁਣ /huṇ/
5	جتھے	ਜਿਥੇ /jithē/, ਜਥੇ /jathē/
6	دسدا	ਦਿਸਦਾ /disdā/, ਦੱਸਦਾ /dassdā/
7	اک	ਅੱਕ /akk/, ਇੱਕ /ikk/
8	جٹ	ਜੱਟ/jaṭṭ/, ਜੁੱਟ /juṭṭ /
9	اس	ਉਸ /us/, ਇਸ /is/
10	اتے	ਅਤੇ /atē/, ਉੱਤੇ /uttē/

Table 1. Ambiguous Shahmukhi Words without Short Vowels

In the written Shahmukhi script, it is not mandatory to put short vowels, called Aerab, below or above the Shahmukhi character to clear its sound leading to potential ambiguous transliteration to Gurmukhi as shown in Table 1. In our findings, Shahmukhi corpus has just 1.66% coverage of short vowels ◌ੁ[ਪ] (0.81415%),

ਿ[ɪ](0.7295%), and ੰ(0.1234%) whereas the equivalent ਿ[ɪ] (4.5462%) and ੁ[ʊ] (1.5844%) in Gurmukhi corpus has 6.13% usage. Hence, it is a big challenge in the process of machine transliteration process to recognize the right word from the written (without diacritic) text because in a situation like this, correct meaning of the word needs to be corroborated from its neighboring context.

Secondly, it is observed that there are multiple possible mappings in Gurmukhi script corresponding to a single character in the Shahmukhi script as shown in Table 2. Moreover, the shown characters of Shahmukhi have vowel-vowel, vowel-consonant and consonant-consonant mapping.

Sr	Char.	Multiple Gurmukhi Mappings
1	و [v]	ਵ [v], ੋ [o], ੌ [ɔ], ੁ [ʊ], ੂ [u], ਓ [o]
2	ی[j]	ਯ [j], ਿ [ɪ], ੇ [e], ੈ[æ], ੀ[i], ਈ [i]
3	ن[n]	ਨ [n], ੰ [ɲ], ਣ [ɳ], ਂ[ɲ]

Table 2. Multiple Mapping into Gurmukhi Script

For example, consider two Shahmukhi words چین /cīn/ and روس /rōs/ having the presence of an ambiguous character ی[i] and و[o] respectively. Our transliteration engine discovers the corresponding word interpretations as ਚੇਨ /cēn/, ਚੀਨ /cīn/, or ਚੈਨ /cain/ and ਰੋਸ /rōs/, or ਰੂਸ /rūs/ respectively. Furthermore, both the problems may coexist in a particular Shahmukhi word, for example, the Shahmukhi word بندی /baṇdī/ which has four different forms ਬਣਦੀ/baṇdī/, ਬੁਣਦੀ/buṇdī/, ਬੰਦੀ/bandī/ or ਬਿੰਦੀ/bindī/ in Gurmukhi script due to ambiguous character ن[n] and missing short vowel. More sample cases are shown in Table 3.

Another variety of word ambiguity mostly found in machine translation systems is where many words have several meanings or sense. The task of word sense disambiguation is to determine which of the sense of an ambiguous word is invoked in a particular use of the word. This is done by looking at the context of the ambiguous word and by exploiting contextual word similarities based on some predefined co-occurrence relations. The various types of disambiguation methods where the source of word similarity was either statistical (Schutze, 1992); (Dagan et al. 1993, 1995); (Karov and Shimon, 1996); (Lin, 1997); or using a manually crafted thesaurus (Resnik, 1992, 1995); (Jiang and Conrath, 1997); is presented in the literature.

Sr	Word with Ambiguous Char.	Possible Gurmukhi Transliteration
1	کھوہ و [v]	ਖੂਹ / khūh/, ਖੋਹ /khōh/
2	پیو ی[j]	ਪਿਓ /piō/, ਪੀਓ /pīō/
3	چین ی[j]	ਚੇਨ /cēn/, ਚੀਨ /cīn/, ਚੈਨ /cain/
4	جاندا ن[n]	ਜਾਂਦਾ /jāndā/, ਜਾਣਦਾ /jāṇdā/
5	سوچنا و [v], ن[n]	ਸੂਚਨਾ /sūcnā/, ਸੋਚਣਾ /sōcaṇā/
6	دین ی[j], ن[n]	ਦੇਣ /dēṇ/, ਦੀਨ /dīn/

Table 3. Shahmukhi Words with Multiple Gurmukhi Mappings

In this paper we have proposed two different algorithms for Shahmukhi word disambiguation. The first algorithm formulates this problem using a state sequence representation as a Hidden Markov Model (HMM). The second approach uses n-gram model in which the joint occurrence of words within a small window of size ± 5 is used.

2 The Level of Ambiguity in Shahmukhi Text

We have performed experiments to evaluate how much word level ambiguity is present in the Shahmukhi text. In order to measure the extent of such ambiguous words in a Shahmukhi corpus we have analyzed the top, most frequent 10,000 words obtained from the Shahmukhi word list that was generated during corpus analysis. The result of this analysis is shown in Table 4.

Sr.	Most Frequent Words	Percentage of Ambiguous words
1	Top 100	20%
2	Top 500	15.8%
3	Top 1,000	11.9%
4	Top 5,000	4.72%
5	Top 10,000	3.6%

Table 4. Extent of Ambiguity in Top 10K words of Shahmukhi Corpus

Observations:

- Most frequent words in Shahmukhi corpus have higher chances of being ambiguous.
- In this test case the maximum amount of ambiguity is 20% which is very high.

- The percentage of ambiguity decreases continuously while moving from most frequent to less frequent words within the list.
- The ambiguous words in Top 10,000 dataset have maximum four interpretations in Gurmukhi script with 2% coverage whereas the amount of three and two Gurmukhi interpretations is 12% and 86% respectively as shown in Figure 2.

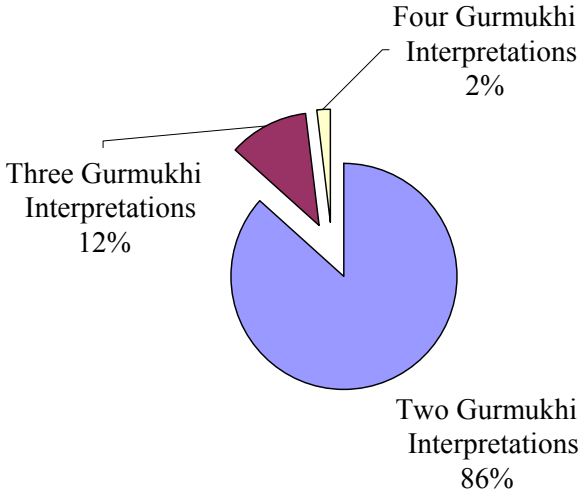

Figure 2. Number of Gurmukhi Interpretations of Ambiguous words in Top 10K Dataset

Additionally, a similar experiment was performed on a Shahmukhi book having a total of 37,620 words. After manual evaluation, we discovered that the extent of ambiguous words in this book was 17.12%. Hence, both the test cases figure out that there is significant percentage of ambiguous words in Shahmukhi text and must be addressed to achieve higher rate of transliteration accuracy.

3 The Approach

At the outset, all we have is the raw corpora for each (Shahmukhi and Gurmukhi) script of Punjabi language. The properties of these corpora are presented in Table 5. The majority of Shahmukhi soft data was found in the form of InPage software files. This soft data was converted to Unicode format using the InPage to Unicode Converter. A corpus based statistical analysis of both the corpora is performed. We have started from scratch and created the required resources in Unicode for both Shahmukhi and Gurmukhi scripts. The size of Gurmukhi training data used for word disambiguation task is shown in Table 6.

N-gram models are used extensively in language modeling and the same is proposed for Shahmukhi word disambiguation using the target script corpus. The N-grams have practical advantages to provide useful likelihood estimations for alternative reading of language corpora. Word similarities that are obtained from N-gram analysis are a combination of syntactical, semantic and contextual similarities those are very suitable in this task of word disambiguation.

Punjabi Script	Corpus Size	Unique words	Text Source
Gurmukhi	7.8 m	1,59,272	Daily and regional news papers, reports, periodicals, magazines, short stories and Punjabi literature books etc.
Shahmukhi	8.5 m	1,79,537	

Table 5. Properties of Shahmukhi and Gurmukhi Corpora

We have proposed two different algorithms for Shahmukhi word disambiguation. The first algorithm formulates this problem using a state sequence representation as a Hidden Markov Model. The second approach uses n-gram model (including the right side context) in which the joint occurrence of words within a small window of size ± 5 is used.

Training Data	Size (Records)
Gurmukhi Word Frequency List	87,962
Gurmukhi Bigram List	265,372
Gurmukhi Trigram List	247,010

Table 6. Training Data Resources

3.1 Word Disambiguation using HMM

Second order HMM is equivalent to n-gram language model with n=3 called trigram language model. One major problem with fixed n models is data sparseness. Therefore, one good idea to smooth n-gram estimates is to use linear interpolation (Jelinek and Mercer, 1980) of n-gram estimates for various n, for example:

$$P(w_n \mid w_{n-1}w_{n-2}) =$$
$$\lambda_1 P_1(w_n \mid w_{n-1}w_{n-2}) + \lambda_2 P_2(w_n \mid w_{n-1}) + \lambda_3 P_3(w_n)$$
$$\text{where } \sum_i \lambda_i = 1 \text{ and } 0 \le \lambda_i \le 1 \qquad (2)$$

The variable n means that we are using trigram, bigram and unigram probabilities together as a linear interpolation. This way we would get some probability of how likely a particular word was, even if our coverage of trigram is sparse.

Now the next question is how to set the parameters λ_i . Thede and Harper, (1999) modeled a second order HMM for part of speech tagging. Rather than using fixed smoothing technique, they have discussed their new method of calculating contextual probabilities using the linear interpolation. This method attaches more weight to triples that occur more often. The formula to estimate contextual probability is:

$$P(\tau_p = w_k \mid \tau_{p-1} = w_j, \tau_{p-2} = w_i) =$$

$$k_3 . \frac{N_3}{C_2} + (1-k_3) k_2 . \frac{N_2}{C_1} + (1-k_3)(1-k_2) . \frac{N_1}{C_0}$$

where $k_3 = \dfrac{\log_2(N_3+1)+1}{\log_2(N_3+1)+2}$; and

$$k_2 = \frac{\log_2(N_2+1)+1}{\log_2(N_2+1)+2} \qquad (3)$$

The equation 2 depends on the following numbers:

N_3 : Frequency of trigram $w_i w_j w_k$ in Gurmukhi corpus

N_2 : Frequency of bigram $w_j w_k$ in Gurmukhi corpus

N_1 : Frequency of unigram w_k in Gurmukhi corpus

C_2 : Number of times bigram $w_i w_j$ occurs in Gurmukhi corpus

C_1 : Number of times unigram w_j occurs in Gurmukhi corpus

C_0 : Total number of words that appears in Gurmukhi corpus

The formulas for k_3 and k_2 are chosen so that the weighting for each element in the equation 2 for *P* changes based on how often that element occurs in the Gurmukhi corpus. After comparing the two equations 1 and 2, we can easily understand that:

$\lambda_1 = k_3$; $\lambda_2 = (1-k_3)k_2$; $\lambda_3 = (1-k_3)(1-k_2)$ and satisfy the condition $\sum_i \lambda_i = 1$; $0 \le \lambda_i \le 1$

We build an HMM with four states for each word pair, one for the basic word pair, and three representing each choice of n-gram model for calculating the next transition. Therefore, as expressed in equation 2 and 3 of second order HMM, there are three ways for w^c ={ਦਸ or ਦੱਸ or ਦਿਸ} to follow $w^a w^b$ (ਹੁਣ ਤੂੰ) and the total probability of seeing w^c next is then the sum of each of the n-gram probabilities that adorn the arcs multiplied by the corresponding parameter $0 \le \lambda_i \le 1$. Correspondingly, the fragment of HMM for ambiguous Gurmukhi word sequence ਹੁਣ ਤੂੰ ਦੱਸ is shown in Figure 3 where the first two consecutive words have two forms {ਹਨ or ਹੁਣ} and {ਤੋਂ or ਤੂੰ} where as the third consecutive Gurmukhi word has three forms like {ਦਸ or ਦੱਸ or ਦਿਸ}.

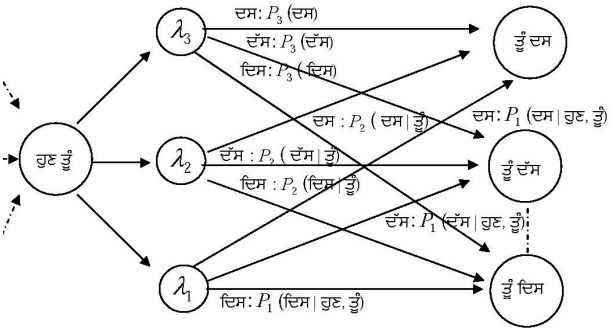

Figure 3. A Fragment of Second Order HMM for ਹੁਣ ਤੂੰ ਦੱਸ

To calculate the best state sequence we have modeled Viterbi Algorithm, which efficiently computes the most likely state sequence.

3.2 Word Disambiguation using ±5 Window Context

This n-gram based algorithm performs word disambiguation using the small window context of size ±5. This context is used to exploit the contextual, semantic and syntactical similarities based on the information captured by an n-gram language model. The structure of our small window context is shown in Figure 4.

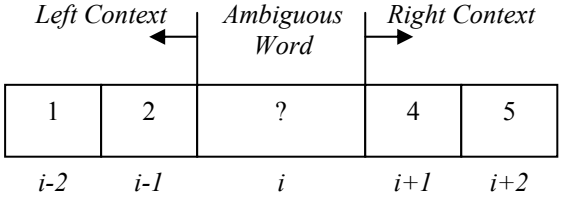

Figure 4. Structure of ±5 Window Context

The disambiguation task starts from the first word of the input sentence and attempts to investigate co-occurrence probabilities of the words present in the left and right of the ambiguous word within the window size. Unlike HMM approach which is based on linear interpolation of n-gram estimates (Jelinek and Mercer, 1980), this algorithm works in a back off fashion as proposed by Katz (1987) in which it first relies

on highest order trigram model to estimate the joint co-occurrence possibility of alternative word interpretations to select the most probable interpretation. For example, consider the following Shahmukhi sentence with three ambiguous words as:

{ਦਸ ਦੱਸ ਦਿਸ} {ਤੋਂ ਤੂੰ} {ਹਨ ਹੁਣ}

The initial ambiguity is {ਹਨ /han/ | ਹੁਣ /huṇ/}, {ਤੋਂ /tōṃ/ | ਤੂੰ /tūṃ/} and {ਦਸ /das/ | ਦੱਸ /dass/ | ਦਿਸ /dis/}. The left and right context probabilities of the first ambiguous word are shown in Table 7.

	Right Context		Left Context	
	Bigram	**Trigram**	**Bigram**	**Trigram**
ਹਨ	P(ਹਨ, ਤੋਂ |ਤੂੰ)	P(ਹਨ, ਤੋਂ |ਤੂੰ, ਦਸ | ਦੱਸ |ਦਿਸ)	P(_ , ਹਨ)	P(_ , _ , ਹਨ)
P=	0.000519	0.0	0.001613	0.001673
ਹੁਣ	P(ਹੁਣ, ਤੋਂ |ਤੂੰ)	P(ਹੁਣ, ਤੋਂ | ਤੂੰ, ਦਸ | ਦੱਸ |ਦਿਸ)	P(_ , ਹੁਣ)	P(_ , _ , ਹੁਣ)
P=	0.015945	**0.037383**	0.063967	**0.064930**

Table 7. Context Window Probabilities for ਹਨ and ਹੁਣ Words

Clearly, word ਹੁਣ is selected because it has higher trigram co-occurrence probability. Now the sentence ambiguity is reduced to ਹੁਣ {ਤੋਂ |ਤੂੰ} {ਦਸ | ਦੱਸ |ਦਿਸ}. The estimation of co-occurrence probability for next ambiguous word is shown in Table 8.

	Right Context	Left Context	
	Bigram	**Bigram**	**Trigram**
ਤੋਂ	P(ਤੋਂ, ਦਸ |ਦੱਸ |ਦਿਸ)	P(ਹੁਣ, ਤੋਂ)	P(_ , ਹੁਣ, ਤੋਂ)
	P=0.000492	P=0.006519	P=0.003341
ਤੂੰ	P(ਤੂੰ, ਦਸ |ਦੱਸ |ਦਿਸ)	P(ਹੁਣ, ਤੂੰ)	P(_ , ਹੁਣ, ਤੂੰ)
	P=0.005019	P=0.009426	**P=0.012454**

Table 8. Context Window Probabilities for ਤੋਂ and ਤੂੰ Words

As expected, word ਤੂੰ is selected by the system using left trigram context and the sentence ambiguity is now reduced to ਹੁਣ ਤੂੰ {ਦਸ ਦੱਸ ਦਿਸ}. The next ambiguity is lying in the last word of the sentence so it has only left context as shown in Table 9. After evaluating the left context co-occurrence for all the three word forms the system found that the valid co-occurrence is P(ਹੁਣ, ਤੂੰ, ਦੱਸ) and on this basis word ਦੱਸ is selected. Finally, the output of the system is ਹੁਣ ਤੂੰ ਦੱਸ as expected.

	Right Context	Left Context	
		Bigram	**Trigram**
ਦਸ	N.A.	P(ਤੂੰ, ਦਸ)	P(ਹੁਣ, ਤੂੰ, ਦਸ)
	-	P=0020768	P=0
ਦੱਸ	N.A.	P(ਤੂੰ, ਦੱਸ)	P(ਹੁਣ, ਤੂੰ, ਦੱਸ)
	-	P=0.003980	**P=0.037383**
ਦਿਸ	N.A.	P(ਤੂੰ, ਦਿਸ)	P(ਹੁਣ, ਤੂੰ, ਦਿਸ)
	-	P=0.000028	P=0

Table 9. Context Window Probabilities for ਦਸ, ਦੱਸ and ਦਿਸ Words

Unlike the above example, there is a situation when the higher order joint co-occurrence is found to be zero in the training corpus. In this situation the proposed algorithm will back off to next lower n-1 gram model.

4 Example

Following is the N-gram and HMM outputs of word disambiguation task for the sample text downloaded from the article available on the web site *http://www.likhari.org*

Input Text

اسیں گلّ تاں کردے ہاں کہ اسیں اپنی ماں بولی نوں اسدا بندا حق دواؤن لئی پرزور ہاں پر ساڈیا اکھاں سامھنے ہی پنجابی نال اسدے گھر وچ ہی نہ انصافی ہو رہی ہے تے اسیں پھر چپّ کر کے ایہہ سبھ ویکھ رہے ہاں بھارت اتے پاکستان دوواں ملکاں ولوں پنجابی لئی ساجھے منچ تے کم کیتا جارہا ہے پچھلے دنیں بمبئی وچ جو کجھ واپریا اس نے ساری دنیاں نوں ہلا کے رکھ دتا اس نال دوواں ملکاں دے رشتے تڑکے ہن پر بدھیجیوی ورگ نوں اک گلّ اپنے ذہن وچ رکھنی چاہیدی ہے کہ سرحداں نے زمین ونڈی ہے زبان نہیں

Romanized:

asīṃ gall tāṃ karadē hāṃ ki asīṃ āpṇī māṃ bōlī nūṃ usdā baṇdā hakk divāuṇ laī purzōr hāṃ par

sāḍīā akkhāṃ sāmhṇē hī pañjābī nāḷ usdē ghar vic hī nā iṃsāfī hō rahī hai tē asīṃ phir cupp kar kē ih sabh vēkh rahē hāṃ bhārat atē pākistān dōvāṃ mulkāṃ vallōṃ pañjābī laī sāñjhē mañc tē kamm kītā jā rihā hai pichlē dinīṃ bambī vic jō kujjh vāpriā us nē sārī dunīāṃ nūṃ hilā kē rakkh dittā is nāḷ dōvāṃ mulkāṃ dē rishtē taṛkē han par buddhījīvī varag nūṃ ikk gall āpaṇē zihan vic rakkhṇī cāhīdī hai ki sarhaddāṃ nē zamīn vaṇḍī hai zabān nahīṃ

N-gram Output:

ਅਸੀਂ **ਗੱਲ** ਤਾਂ ਕਰਦੇ ਹਾਂ ਕਿ ਅਸੀਂ ਆਪਣੀ ਮਾਂ ਬੋਲੀ **ਨੂੰ** ਉਸਦਾ **ਬਣਦਾ** ਹੱਕ ਦਿਵਾਉਣ ਲਈ ਪੁਰਜ਼ੋਰ ਹਾਂ **ਪਰ ਸਾਡਿਆ** ਅੱਖਾਂ ਸਾਮ੍ਹਣੇ ਹੀ ਪੰਜਾਬੀ ਨਾਲ਼ ਉਸਦੇ ਘਰ ਵਿਚ ਹੀ ਨਾ ਇੰਸਾਫ਼ੀ ਹੋ ਰਹੀ ਹੈ ਤੇ ਅਸੀਂ ਫਿਰ ਚੁੱਪ ਕਰ ਕੇ ਇਹ ਸਭ ਵੇਖ ਰਹੇ ਹਾਂ ਭਾਰਤ **ਅਤੇ** ਪਾਕਿਸਤਾਨ ਦੋਵਾਂ ਮੁਲਕਾਂ ਵੱਲੋਂ ਪੰਜਾਬੀ ਲਈ ਸਾਂਝੇ ਮੰਚ ਤੇ ਕੰਮ ਕੀਤਾ ਜਾ ਰਿਹਾ ਹੈ ਪਿਛਲੇ ਦਿਨੀਂ ਬੰਬਈ ਵਿਚ ਜੋ ਕੁੱਝ ਵਾਪਰਿਆ **ਉਸ** ਨੇ ਸਾਰੀ ਦੁਨੀਆਂ **ਨੂੰ** ਹਿਲਾ ਕੇ **ਰੱਖ** ਦਿੱਤਾ **ਇਸ** ਨਾਲ਼ ਦੋਵਾਂ ਮੁਲਕਾਂ ਦੇ ਰਿਸ਼ਤੇ ਤੜਕੇ **ਹਨ ਪਰ** ਬੁੱਧੀਜੀਵੀ ਵਰਗ **ਨੂੰ ਇਕ ਗੱਲ** ਆਪਣੇ ਜ਼ਿਹਨ ਵਿਚ ਰੱਖਣੀ ਚਾਹੀਦੀ ਹੈ ਕਿ ਸਰਹੱਦਾਂ ਨੇ ਜ਼ਮੀਨ ਵੰਡੀ ਹੈ ਜ਼ਬਾਨ ਨਹੀਂ

Ambiguous words (Total =15 i.e. 14.285%)

{ਗੱਲ ਗਿੱਲ ਗੁੱਲ ਗੁਲ}{ਨੂੰ ਨੈਂ}{ਬਣਦਾ ਬੰਦਾ}{ਪਰ ਪ੍ਰ ਪੁਰ}{ਸਾਡਿਆ ਸਾਡੀਆ}{ਅਤੇ ਉੱਤੇ}{ਉਸ ਇਸ}{ਨੂੰ ਨੈਂ}{ਰੱਖ ਰੁੱਖ}{ਇਸ ਉਸ ਐਸ}{ਹਨ ਹੁਣ}{ਪਰ ਪ੍ਰ ਪੁਰ}{ਨੂੰ ਨੈਂ}{ਇਕ ਅੱਕ ਇੱਕ}{ਗੱਲ ਗਿੱਲ ਗੁੱਲ ਗੁਲ}

2nd Order HMM Output:

ਅਸੀਂ **ਗੱਲ** ਤਾਂ ਕਰਦੇ ਹਾਂ ਕਿ ਅਸੀਂ ਆਪਣੀ ਮਾਂ ਬੋਲੀ **ਨੂੰ** ਉਸਦਾ **ਬਣਦਾ** ਹੱਕ ਦਿਵਾਉਣ ਲਈ ਪੁਰਜ਼ੋਰ ਹਾਂ **ਪਰ ਸਾਡੀਆ** ਅੱਖਾਂ ਸਾਮ੍ਹਣੇ ਹੀ ਪੰਜਾਬੀ ਨਾਲ਼ ਉਸਦੇ ਘਰ ਵਿਚ ਹੀ ਨਾ ਇੰਸਾਫ਼ੀ ਹੋ ਰਹੀ ਹੈ ਤੇ ਅਸੀਂ ਫਿਰ ਚੁੱਪ ਕਰ ਕੇ ਇਹ ਸਭ ਵੇਖ ਰਹੇ ਹਾਂ ਭਾਰਤ **ਅਤੇ** ਪਾਕਿਸਤਾਨ ਦੋਵਾਂ ਮੁਲਕਾਂ ਵੱਲੋਂ ਪੰਜਾਬੀ ਲਈ ਸਾਂਝੇ ਮੰਚ ਤੇ ਕੰਮ ਕੀਤਾ ਜਾ ਰਿਹਾ ਹੈ ਪਿਛਲੇ ਦਿਨੀਂ ਬੰਬਈ ਵਿਚ ਜੋ ਕੁੱਝ ਵਾਪਰਿਆ **ਉਸ** ਨੇ ਸਾਰੀ ਦੁਨੀਆਂ **ਨੂੰ** ਹਿਲਾ ਕੇ **ਰੱਖ** ਦਿੱਤਾ **ਇਸ** ਨਾਲ਼ ਦੋਵਾਂ ਮੁਲਕਾਂ ਦੇ ਰਿਸ਼ਤੇ ਤੜਕੇ **ਹਨ ਪਰ** ਬੁੱਧੀਜੀਵੀ ਵਰਗ **ਨੂੰ ਇੱਕ ਗੱਲ** ਆਪਣੇ ਜ਼ਿਹਨ ਵਿਚ ਰੱਖਣੀ ਚਾਹੀਦੀ ਹੈ ਕਿ ਸਰਹੱਦਾਂ ਨੇ ਜ਼ਮੀਨ ਵੰਡੀ ਹੈ ਜ਼ਬਾਨ ਨਹੀਂ

This sample input text has 105 words in total and around 14.28% ambiguity at word level. While processing, the disambiguation task identified that there are fifteen (bold face) words that are ambiguous, i.e. having two, three, and four interpretations in Gurmukhi script. The disambiguation results of this sample input show that out of fifteen ambiguous words fourteen have been correctly disambiguated by both the N-gram and HMM algorithms whereas only one wrong word ਸਾਡਿਆ /sāḍiā/ is mistakenly chosen by N-gram approach that has correctly recognized as ਸਾਡੀਆ /sāḍīā/ by the HMM algorithm.

5 Experiments and Results

The natural sources of Shahmukhi text are very limited. With this limitation we have identified the available online and offline sources and three different test sets are taken from different domains as shown in Table 10. After manual evaluation, the word disambiguation results on the three datasets are given in Table 11. The overall 13.85% word ambiguity corresponding to all datasets has a significant value. The upper bound contribution is from Set-1(book) having a highest percentage 17.12% of word ambiguity and the corresponding performance of two different disambiguation tasks is also highest.

Test Data	Word Size	Source
Set-1	37,620	*Book*
Set-2	39,714	*www.likhari.org*
Set-3	46678	*www.wichaar.com*
Total	**1,24,012**	

Table 10. Description of the Test Data

We have evaluated both HMM and N-gram algorithms on these datasets and the results of this experiment have shown that the accuracy of N-grams and HMM based algorithms is 92.81% and 93.77% respectively. Hence, the HMM based approach has outperformed marginally.

Test Data	Word Ambiguity	N-gram size ± 5	2nd order HMM
Set-1 (book)	17.121%	95.358%	95.870%
Set-2 (likhari.org)	12.587%	91.189%	91.629%
Set-3 (wichaar.com)	11.85%	91.892%	93.822%
Total	**13.85%**	**92.813%**	**93.773%**

Table 11. Word Disambiguation Result

The accuracy of both algorithms is more that 92%, indicating there is still room for improvement. A comparative analysis of both outputs is performed. We found that there are cases when both HMM and N-gram based methods individually outperform as shown in Table 12 row 1

& 2 and row 3 & 4 respectively. However, there are various cases in which both approaches fail to disambiguate either partially or fully as shown in row 5 & 6 of Table 12. It is observed that due to lack of training data both the proposed approaches have failed to distinguish correctly like ਅਤੇ /atē/ or ਉੱਤੇ /uttē/ as shown in 5th row of Table 12. Similarly, in some other cases system fails to predict name entity abbreviations as shown in 6th row of Table 12.

We can produce better results in the future by increasing the size of the training corpus and by exploiting contextual word similarities based on some predefined co-occurrence relations.

Sr.	N-gram Output	Word Ambiguity Correct = ✓	2nd order HMM Output
1	ਕਈ ਵਾਰ ਲੋਕਾਂ ਵਿਚੋਂ ਕੁੱਝ ਲੋਕ ਵੀ **ਇਕ** ਕੇ ਜ਼ੁਲਮ ਦਾ ਟਾਕਰਾ ਕਰਨ ਲਈ ਹਥਿਆਰ ਚੱਕ ਲੈਂਦੇ ਹਨ ਪਰ ਸੇਧ ਅਤੇ ਅਨੁਸ਼ਾਸਨ ਦੀ ਘਾਟ ਕਾਰਨ ਅਪਣਾ ਹੀ ਨੁਕਸਾਨ ਕਰਵਾ ਬੈਠਦੇ ਹਨ kaī vār lōkāṃ vicōṃ kujjh lōk vī ik kē zulam dā ṭākrā karan laī hathiār cakk laindē han par sēdh atē anushāsan dī ghāṭ kāran apṇā hī nuksān karavā baiṭhdē han	{ਇਕ **ਅੱਕ** ✓ ਇੱਕ} {ਹਨ ਹੁਣ} {ਪਰ ਪ੍ ਪੁਰ} {ਅਤੇ ਉੱਤੇ} {ਹਨ ਹੁਣ} {ik akk ikk} {han huṇ} {par pr pur} {atē uttē} {han huṇ}	ਕਈ ਵਾਰ ਲੋਕਾਂ ਵਿਚੋਂ ਕੁੱਝ ਲੋਕ ਵੀ **ਅੱਕ** ਕੇ ਜ਼ੁਲਮ ਦਾ ਟਾਕਰਾ ਕਰਨ ਲਈ ਹਥਿਆਰ ਚੱਕ ਲੈਂਦੇ ਹਨ ਪਰ ਸੇਧ ਅਤੇ ਅਨੁਸ਼ਾਸਨ ਦੀ ਘਾਟ ਕਾਰਨ ਅਪਣਾ ਹੀ ਨੁਕਸਾਨ ਕਰਵਾ ਬੈਠਦੇ ਹਨ kaī vār lōkāṃ vicōṃ kujjh lōk vī akk kē zulam dā ṭākrā karan laī hathiār cakk laindē han par sēdh atē anushāsan dī ghāṭ kāran apṇā hī nuksān karavā baiṭhdē han
2	ਆਪਣੇ ਘਰਦਿਆਂ **ਤੂੰ** ਵਿਆਹ ਦੀ ਆਸ ਲਾਹ ਛੱਡ āpaṇē ghardiāṃ tūṃ viāh dī ās lāh chaḍḍ	{**ਤੂੰ ਤੋਂ** ✓ } {tūṃ tōṃ}	ਆਪਣੇ ਘਰਦਿਆਂ **ਤੋਂ** ਵਿਆਹ ਦੀ ਆਸ ਲਾਹ ਛੱਡ āpaṇē ghardiāṃ tōṃ viāh dī ās lāh chaḍḍ
3	ਬਚਪਨ ਵਿਚ ਅਸੀਂ ਉਨ੍ਹਾਂ ਕੋਲ਼ **ਜਾਂਦੇ** ਹੁੰਦੇ ਸੀ bacpan vic asīṃ unhāṃ kōḷ jāndē hundē sī	{**ਜਾਂਦੇ** ✓ ਜਾਣਦੇ} {jāndē jāṇdē}	ਬਚਪਨ ਵਿਚ ਅਸੀਂ ਉਨ੍ਹਾਂ ਕੋਲ਼ **ਜਾਣਦੇ** ਹੁੰਦੇ ਸੀ bacpan vic asīṃ unhāṃ kōḷ jāṇdē hundē sī
4	ਸਾਨੂੰ ਪੁੱਤਾਂ ਤੋਂ ਸਪਤ ਬਣਨ ਲਈ ਜੀਵਨ ਸੇਧ ਗੁਰੂ ਗ੍ਰੰਥ ਸਾਹਿਬ ਦੇ ਵਿਚ ਦਰਜ ਬਾਣੀ **ਤੋਂ** ਹੀ ਮਿਲ ਸਕਦੀ ਹੈ sānūṃ puttāṃ tōṃ sapat baṇan laī jīvan sēdh gurū granth sāhib dē vic daraj bāṇī tōṃ hī mil sakadī hai	{ਤੋਂ ਤੂੰ} {**ਤੋਂ** ✓ ਤੂੰ} {ਮਿਲ ਮੱਲ ਮਿੱਲ ਮੁੱਲ} {tōṃ tūṃ} {tōṃ tūṃ} {mil mall mill mull}	ਸਾਨੂੰ ਪੁੱਤਾਂ ਤੋਂ ਸਪਤ ਬਣਨ ਲਈ ਜੀਵਨ ਸੇਧ ਗੁਰੂ ਗ੍ਰੰਥ ਸਾਹਿਬ ਦੇ ਵਿਚ ਦਰਜ ਬਾਣੀ **ਤੂੰ** ਹੀ ਮਿਲ ਸਕਦੀ ਹੈ sānūṃ puttāṃ tōṃ sapat baṇan laī jīvan sēdh gurū granth sāhib dē vic daraj bāṇī tūṃ hī mil sakadī hai
5	ਮੇਰੇ **ਉੱਤੇ** ਅਨੁਪ੍ਰੀਤ ਦੇ ਹੈਮਿਲਟਨ ਰਹਿਣ ਕਰ ਕੇ ਜਸਜੀਤ ਵੀ ਸਾਡੇ ਪਾਸ ਆ ਗਈ mērē uttē anuprīt dē haimilṭan rahiṇ kar kē jasjīt vī sāḍē pās ā gaī	{**ਅਤੇ** ✓ ਉੱਤੇ} {atē uttē}	ਮੇਰੇ **ਉੱਤੇ** ਅਨੁਪ੍ਰੀਤ ਦੇ ਹੈਮਿਲਟਨ ਰਹਿਣ ਕਰ ਕੇ ਜਸਜੀਤ ਵੀ ਸਾਡੇ ਪਾਸ ਆ ਗਈ mērē uttē anuprīt dē haimilṭan rahiṇ kar kē jasjīt vī sāḍē pās ā gaī
6	ਪ੍ਰੋਫ਼ੈਸਰ **ਏਸ** ਐਨ ਮਿਸ਼ਰਾ prōfaisar ēs ain mishrā	{ਏਸ ਇਸ **ਐਸ** ✓ } {ਐਨ ਇਨ} {ēs is ais} {ain in}	ਪ੍ਰੋਫ਼ੈਸਰ **ਇਸ** ਐਨ ਮਿਸ਼ਰਾ prōfaisar is ain mishrā

Table 12. Sample Failure Cases

References

Dekang Lin. 1997. Using syntactic dependency as local context to resolve word sense ambiguity. In *Proceedings of 35th Annual Meeting of the ACL and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain, 64-71.

Frederick Jelinek and Robert L. Mercer. 1980. Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice*. Amsterdam, The Netherlands: North-Holland.

Harbhajan Singh. 1997. *Medieval Indian Literature: An Anthology*. Paniker K. Ayyappa, (Ed.) Sahitya Akademi Publication, volume 2, 417-452.

Hinrich Schütze. 1992. Dimensions of meaning. In *Proceedings of Supercomputing '92*. Minneapolis, MN, 787-796.

Ido Dagan, Shaul Marcus and Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics (ACL '93)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 164-171. doi=10.3115/981574.981596

Ido Dagan, Shaul Marcus and Shaul Markovitch. 1995. Contextual word similarity and estimation from sparse data. *Computer Speech and Language*, 9:123-152.

Jay J. Jiang. David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING)*. Taiwan, 1-15.

Lawrence R. Rabiner. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257-285.

Nadir Durrani and Sarmad Hussain. 2010. Urdu word segmentation. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, California, 528-536.

Philip Resnik. 1992. WordNet and distributional analysis: A class-based approach to lexical discovery. In *Proceedings of AAAI Workshop on Statistically-based Natural Language Processing Techniques*. Menlo Park, California, 56-64.

Philip Resnik. 1995. Disambiguating noun groupings with respect to WordNet senses. In *Proceedings of the Third Workshop on Very Large Corpora*. Cambridge, 54-68

Ralph Grishman and John Sterling. 1993. Smoothing of automatically generated selectional constraints. In *Proceedings of DARPA Conference on Human Language Technology*. San Francisco, California, 254-259.

Ralph Grishman, Lynette Hirschman and Ngo Thanh Nhan. 1986. Discovery procedures for sublanguage selectional patterns: initial experiments. *Computational Linguistics*, 12(3):205-215.

Sant S. Sekhon. 1996. *A History of Panjabi Literature*, Publication Bureau, Punjabi University, Patiala, volume 1 & 2, Punjab, India.

Scott M. Thede and Mary P. Harper. 1999. A second-order Hidden Markov Model for part-of-speech tagging. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics (ACL '99)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 175-182. doi=10.3115/1034678.1034712

Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP, 35(3):400-401.

Ute Essen and Volker Steinbiss. 1992. Cooccurrence smoothing for stochastic language modeling. In *Proceedings of the 1992 IEEE international conference on Acoustics, speech and signal processing - Volume 1 (ICASSP'92)*, IEEE Computer Society, Washington, DC, USA, 161-164.

Yael Karov and Shimon Edelman. 1996. Learning similarity-based word sense disambiguation from sparse data. In *Proceedings of the Fourth Workshop on Very Large Corpora*. Copenhagen, Denmark, 42-55.