Bi-Gram based Probabilistic Language Model for Template Messaging

Rina Damdoo
Department of CSE
RCOEM
Nagpur, MS, INDIA

ABSTRACT
This work reports the benefits of Statistical Machine Translation (SMT) in template messaging domain. SMT has become an actual and practical technology due to significant increment in both the computational power and storage capacity of computers and the availability of large volumes of bilingual data. Through SMT a sentences written with misspelled words, short forms and chatting slang can be corrected. The problem of machine translation is to automatically produce a target-language (e.g., Long form English) sentence from a given source-language (e.g., Short form message) sentence. SMS Lingo is a language used by youngsters for instant messaging or for chatting on social networking websites called chatting slang. Such terms often originate with the purpose of saving keystrokes.

This work presents a pioneering step in designing Bi-Gram based back-off decoder for template messages. Among the different machine translation approaches, the Probabilistic N-gram-based system has proved to be comparable with the state-of-the-art phrase-based systems. In N-gram Language Model N words are used to find the context of a word. In this work, Bi-gram LM is used. First LM is trained with bi-lingual parallel word aligned corpus to get Probability Distribution Tables (Bi-gram PDT and Uni-gram PDT). Back-off decoder along with these PDTs is then employed to translate template messages into full form text.

Idea behind this work is to deal with text normalization as a translation task with the Bi-gram-based system. The main goal behind this project is to analyze the improvement in efficiency of Language Model as the size of bilingual corpus increases. This work will help researchers as a lead way in the field of N-Gram Probabilistic Machine Translation and Human Computer Interaction.

This work will help users to combine multiple languages with larger vocabulary and is a useful tool for small devices like mobile phones. It is also a time saver for those who cannot operate the keys efficiently. Machine learning and translation systems, dictionary and textbook preparations, patent and reference searches and various information retrieval systems are the main applications of the work.

General Terms
Statistical Machine Translation, Probabilistic Language Model, Back-off decoder, N-grams, Probability distribution Table

Keywords
SMS Lingo, Bi-grams, Template messaging

1. INTRODUCTION
Internet users have popularized, Internet slang (Internet short-hand, netspeak or chatspeak), a type of slang that have benefited in many cases. SMS Lingo is a language used by young generation for instant messaging. Such terms often originate with the purpose of saving keystrokes. Many people use the same abbreviations in texting and instant messaging (u mean you), and social networking websites. Acronyms, keyboard symbols and shortened words are often used as methods of abbreviation in Internet slang. New dialects of slang, such as leet or Lolspeak develop as in-group memes rather than time savers.

Secondly, over past few years social networks, chat rooms and forums have become the most important websites for users to share information about their life, work and interests. This new way of communication has evolved in such a way that they all share a casual common language. The users write on these sites as if they were writing SMS messages on their mobile phones, without paying attention to correct spelling or moreover, using user-created abbreviations for common phrases, e.g. "how are you?" is commonly written as "h r u?". For these reasons, existing natural language processing tools cannot process the generated content found on the websites. Simple tools like dictionaries are not entirely satisfying, because the same abbreviation may have several expansions with different meanings ("2" could either mean "too", "to" or "two") and a context analysis evaluation should be made to choose the right definition. A machine translation system may address this challenge because it considers both the translation model, which would offer the different meanings for the same abbreviation or misspelled word, and the context analysis, which would consider the current context to choose the best translation. Figure 1shows, example text messaging of two persons on mobile phones. Both users are texting short message, but the end user is able to see the long form text message with increased readability.

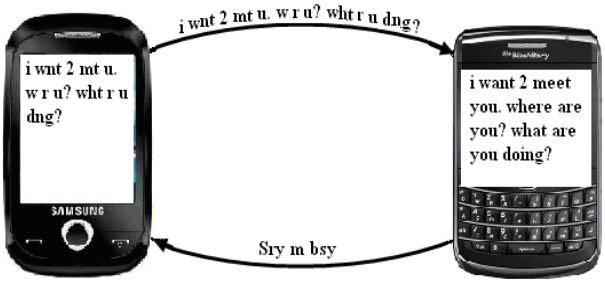

Fig 1: Example text messaging of two persons on mobile phones

Category	Definition	Example
Deletion	**Deletions. No substitutions, insertions or order change.**	**"pctr" (picture)**
Clipping	Entire syllable(s) deleted	"cafet"(Cafeteria)
Location	Silent 'e', Initial 'h' or -ing 'g' removed	"goin" (going)
General	From a single word, not a Clipping or Location.	"pk" (pick)
Initialization	First letter of each word in a multi-word phrase	"gn" (good night)
Substitution	**Using character combinations with similar sounds or looks**	**"2de" (today)**
As Sound	Reading the substitution as a word	"1ce" (once)
As Character	Pronouncing character's names individually	"ne"(any)
Symbolic	Use of symbol(s) that "look like" the character	"$ory" (sory)
Combination	**Use of two or more of the above in a single word.**	**"2mro" (Tomorrow)**

Table 1: Categories of Texting Abbreviations [1]

Given a short form Template message, objective of the work is to translate it into full form template message without extra spaces, using Bi-Gram Language Model and Back-off Decoder.

For Example:

Input → H r u ? I m fin. V wil , mt tom!

Output → How are you? I am fine. We will meet tomorrow!

While messaging, one tries to type maximum information in single SMS. Secondly, young generation does not pay attention to grammar. Instead of writing "I am waiting", "am waiting" or "I waiting" or "me waiting" is used. Thirdly, the consequence of using this casual language is, word based translation model does fail if a person uses same abbreviation for more than one word. One writes same abbreviation "wh", sometimes for "what", sometimes for "where", sometimes for "why", sometimes for "who", so to get the context clearer the earlier and/or later words also must be considered. In short a context analysis evaluation should be made to choose the right definition [1, 2, 3, 6]. Table 1 gives some sample abbreviations with their expanded definitions. Table 2 lists some sample abbreviations with their multiple expanded definitions. Dictionary and Textbook Preparations, Patent and Reference Searches, Information Retrieval Systems, Chat-Rooms for Chat-Speak Style Communications, Mobile Phone Messaging and Vocabulary Improvement for Kids' Self-Learning are different applications of the work.

Table 2. Sample Abbreviations with their Multiple Expanded Definitions

Abbreviation	Expanded Definitions
lt	Let, Late
the	The, There, Their
n	In, And
me	Me, May
wer	Were, Wear, Where
dr	Dear, Deer, Doctor
w	Where, What, Why, Who, We

2. SYSTEM MODEL

Short form SMS collection involves collection of sample SMS from different data sources. A fixed set of full form messages is provided to all the data sources and corresponding short form messages are collected. All these messages are then preprocessed by separating the sentences from each other, removing extra spaces and removing punctuation marks. After statistical study of data collected, called corpus, probability of occurrence of a word or sequence of words is evaluated in N-Gram Probability Estimation phase giving Probability Distribution Tables (PDT). A PDT is a three column table containing text to be translated, its translation text and probability of this translation. N-Gram Decoder uses these tables to find the most probable translation for the input text. Figure 2 shows system model for the present work.

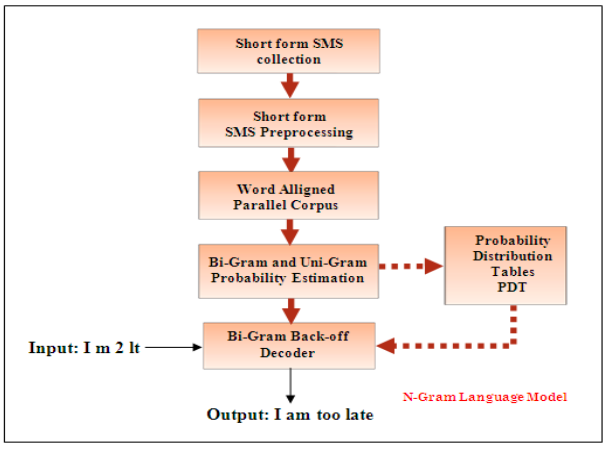

Fig 2: System Model

2.1 Bi-Gram Based Statistical Machine Translation System

In Table 3, bi-lingual corpora can be seen. This corpus is first preprocessed to replace punctuation marks by spaces, remove extra spaces and all those characters which can't be processed by the further modules. Probabilities for the Uni-Grams and Bi-Grams are then calculated to get the Uni-Gram and Bi-Gram PDTs. These PDTs are then used by the decoder to translate a short template messages into long form template messages.

Role of Bi-Gram Language Model in word context disambiguation in machine translation can be explained as follows. Uni-Gram 'nt' in the present language model can be an abbreviation for

nt → note

nt → night

nt → not

By what it must be replaced, is decided by the decoder after looking at its preceding word. If it is 'good', 'nt' must be 'night', if it is 'do','why' or 'am', it must be 'not' and if it is 'please', chances of 'nt' representing 'note' are more than 'night' or 'not'.

Pl nt → Please note

Gd nt → Good night

Y nt → Why not

SMT has two major components [3, 13]:

- A phrase-table/probability distribution table
- A decoder

2.1.1 Probability Distribution Table

A Probability Distribution Table (PDT) or Phrase Table captures all the possible translations of each source phrase. The probability of a target phrase is:

$$P(e \mid f) = \frac{freq(e, f)}{freq(f)} \qquad (2.1)$$

$$P(and \mid n) = \frac{freq(and, n)}{freq(n)} \qquad (2.2)$$

Here, counts are gathered from a word-aligned parallel corpus. ***e*** is a target phrase and ***f*** is a source phrase. It can be useful to model ***P(f | e). P(and | n)*** gives the probability of ***'n'*** getting translated to ***'and'.***

Here if phrase e consists of a string of n words:

e= w1w2w3… wn

= I am free and need your directions

The Bi-Gram model approximates the probability of a word given the previous word ***P(wn|wn-1).***

***P*(I am)= *P*(I) *P*(am| I)**

Most important task in training phase is generation of probability distribution tables. In the given training corpus, Table 3, Uni-Gram 'n' occurs 8 times, out of which only 2 times it represents 'in' and 6 times it represents 'and'. So, the probability of 'n' getting translated to 'in', using equation 2.1 is given as:

$$P(in \mid n) = \frac{freq(in, n)}{freq(n)} = 2 / 8 = 0.25$$

Probability of 'n' getting translated to 'and' is given as:

$$P(and \mid n) = \frac{freq(and, n)}{freq(n)} = 6 / 8 = 0.75$$

Similarly, Bi-Gram 'w r' occurs 6 times, out of which only 2 times it represents 'where are' and 4 times it represents 'what are'. So, the probability of 'w r' getting translated to 'where are' is given as:

$$P(where\ are \mid w\ r) = \frac{freq(where\ are, w\ r)}{freq(w\ r)} = 2 / 6 = 0.33$$

Probability of 'w r' getting translated to 'what are' is given as:

$$P(what\ are \mid w\ r) = \frac{freq(what\ are, w\ r)}{freq(w\ r)} = 4 / 6 = 0.66$$

Table 4 and 5 show a sample Uni-Gram PDT and Bi-Gram PDT for the corpus in Table 3, respectively. PDT is a three column table of Uni-Gram (or Bi-Gram) f, translation e and corresponding probability P(e|f) of f getting translated to e.

System generates Uni-Grams and Bi-Grams probability distribution tables by using following steps:

- Extraction of Uni-Gram (or Bi-Gram) from parallel corpus
- Keep count of occurrence of Uni-Gram (or Bi-Gram) short form
- Record the corresponding long form for each short form Uni-Gram (or Bi-Gram)
- Calculate the probability of short form Uni-Gram (or Bi-Gram)
- Keep the short form, long form and probability value in Uni-Gram (or Bi-Gram) Probability Distribution Table.

These tables are used by Bi-Gram Decoder to translate input text to its most probable output text. In this work Back-off decoder is employed.

2.1.2 N-Gram Decoder

A decoder actually searches for the best target translation given a source sentence. It uses the PDT and language models [7]. The question here in the sentence 'I m fre n ned ur directns' n should be translated to 'and' or 'in' as in word aligned parallel corpus 'n' is representing both. This can be decided by decoder after searching in Bi-Gram PDT and 'n' is translated to 'and' if

***P*(free and | fre n) > *P*(free in | fre n)**

Table 3: Word-Aligned Parallel Corpus

Target language (e)	Source language (f)
sorry busy at moment. I am in meeting, please call me later, I cant pick your phone now so sorry. see you later. I am too late so sorry but I will be there soon, because I need your help. where are you? I want to meet you. what are you doing? I reached before you! are you busy? can we talk? I am free and need your directions. I will call you later. Please help, meet you at four. what are the plans for today? Let us go out and have party.	sry bsy @ momnt. I m n mtng, pl cal me ltr, I cnt pik ur ph now so sry. C u ltr. I m 2 lat so sry but I wl b ther sun, bcas I ned ur hlp. w r u? I wnt 2 mt u. w r u dong? I rchd bfor u! r u bz? cn v tlk? I m fre n ned ur directns. I wil cal u ltr. Pls hlp, met u @ 4. w r th plns 4 2day? lt us go out n hv prty.
	sry bsy @ momnt. I m n mtng, pl cal me ltr, I cnt pik ur ph now so sory. c u ltr. I m 2 lt so sry bt I wil b the sun, coz I ned ur hlp. w r u? I wnt 2 mt u. w r u dng? I rechd b4 u! r u bsy? can v tlk? I m fre n ned ur directns. I wil cal u ltr. Pl hlp, mt u @ four. w r the plans 4 2de? let us go out n hv prty.
	srry bg at mmnt. I m in meet, pls cl me latr, I cnt pck ur phn nw so srry. c u latr. I m 2 late so srry bt I wl b thr soon, coz I nd ur hlp. whr r u? I wnt 2 mt u. wht r u doin? I rchd bfr u! r u bg? cn v tlk? I am free n nd ur dirctions. I wl cl u latr. Pls hlp, mt u at 4. wht r d plns fr 2day? Let us go out n hv prty.
	sry bsy at moment. I am in meeting, pls call me later, I cant pick your phn now so sry. see you ltr. I am too late so sry but I will be thr soon, bcaz I need your hlp. whr are u? I wnt to meet u. wht are u dng? I reached bfr u! are u busy? can we talk? I am free and need your directions. I will call u ltr. Pls hlp, meet u at four. wht are the plans for tdy? Let us go out and hve party.
	sory bzy @ momnt. I m n mtng, pls cal me latr, I cnt pck ur ph now so sory. c u latr. I m 2 lat so sry bt I wil b ther sun, bcoz I ned ur hlp. w r u? I wnt 2 mt u. w r u dng? I rechd b4 u! r u bzy? cn we tlk? I m fre n ned ur dirctns. I wl cal u latr. Pls hlp, mt u @ 4. wh r the plans 4 2de? Let us go out n hv prty.

Table 4: Sample Uni-Gram PDT for the Corpus in Table 3

Short form Uni-gram (f)	Long form Translation (e)	P(e/f) =freq(e,f)/freq(f)
m	**am**	**1.00**
w	**where**	**0.60**
w	**what**	**0.40**
r	**are**	**1.00**
2	**to**	**0.78**
2	**too**	**0.22**
wnt	**want**	**1.0**

Table 5: Sample Bi-Gram PDT for the Corpus in Table 3

Short form Bi-gram (f)	Long form Translation (e)	P(e/f) =freq(e,f)/freq(f)
m 2	**am too**	**1.00**
w r	**where are**	**0.33**
w r	**what are**	**0.66**
r u	**are you**	**1.00**
2 atnd	**2 attend**	**1.00**
2 lt	**too late**	**1.00**
wnt	**want to**	**1.00**

2.1.3 Back-Off Decoder

A Due to the lack of enough amount of training corpus, word probability distribution [9, 13] is misrepresented. In a back-off model if order of word pair is not found within the definite context in training corpus the higher N-gram tagger is backed off to the lower N-Gram tagger. For example Tri-gram will back off to Bi-gram, Bi-gram will back off to Uni-gram and in turn Uni-gram also can back off performing the default task [3, 12].

3. SYSTEM IMPLEMENTATION

The work is divided mainly into four phases as shown in Figure 3:

- **Preprocessing:** PDT generation
- **Training:** PDT generation
- **Testing:** Implementation of decoder
- **Evaluation:** Evaluation of Bi-Gram Language Model

Step 1: Preprocessing

In development and testing process, in first phase data for the project work is collected from 5 persons, which is each of, 100 template messages and 1622 words. Data collected from the sources is preprocessed and a word aligned parallel corpus is created. Word aligned parallel corpus is a bi-lingual large collection of parallel texts. Preprocessing of the text involves following stages and is required in all the stages of implementation.

- Separation of sentences by searching punctuation marks **?**, **!** and **.**
- Replacement of punctuation marks **,** , **;**, **:** by spaces
- Removal of extra spaces

Step 2: Training of Language Model

Out of the source data, data from 4 persons is used to train the Language Model and data from fifth person is retained for testing purpose. First 4 unmerged rows in this table show the training data and fifth row shows the testing data. The longform text is saved in file named full.txt and four training samples of shortforms are saved in train1.txt, train2.txt, train3.txt and train4.txt. Training samples are word aligned parallel with full.txt one by one and the model is trained.

Step 3: Testing of Language Model

Out of the data from sources, data from fifth persons is used to test the Language. Fifth row shows the testing data. The testing sample of shortforms is saved in test.txt using PDTs and back-off decoder. The translated text is saved in decode.txt. This file must be then compared with the expected output file.

Step 4: Performance Evaluation of Language Model

Out of source data provided by 5 persons, data from 4 persons is used during training and data from fifth person is used during testing.

4. RESULTS AND DISCUSSION

This project produces correct translations for the seen words and unseen words are output without any alteration. The results are also user dependent. If one uses some other short form than the data provided while training the model expected results are changed. Also for some Bi-grams like 'w r' the results depend on the data provided for training. For example 'where are you' and 'what are you doing', these two sentences are there in the corpus and both can be represented by Bi-gram 'w r' so there are chances of getting output 'what are you' instead of 'where are you', because probability of 'w r' as 'where are' is 0.3333 and probability of 'w r' as 'what are' is 0.5555.

As in this work, word aligned parallel corpus is used contraction 'btw' can't represent phrase 'by the way'. Lack of training data, changes in the human domain where the system is used, or differences in morphology, length of template message affect the performance of the project.

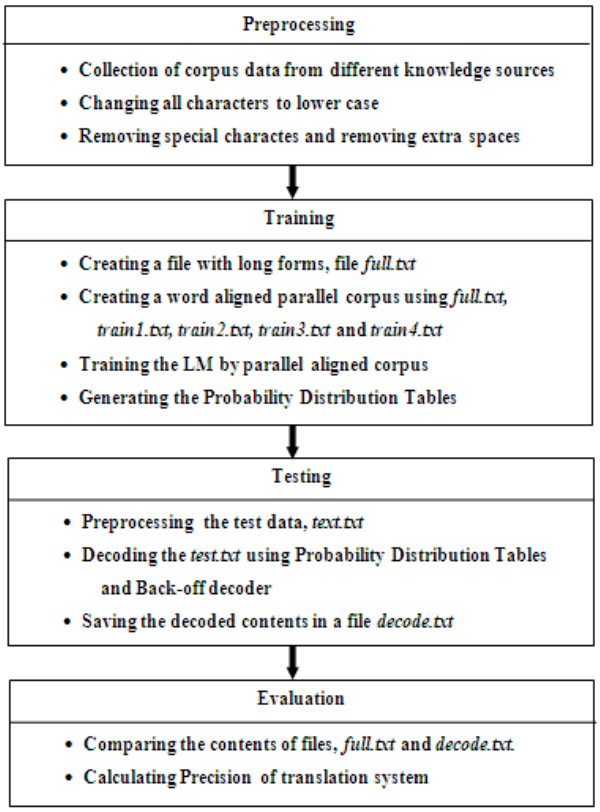

Fig 3: Implementation Stages

4.1 Uni-Gram PDT for the present work

Table 6 shows the Resultant Uni-Gram Probability Distribution Table. Here only 6 selected entries out of 1084 entries of Uni-Gram PDT are shown.

Table 6. Uni-gram Probability Distribution Table

Short form Uni-gram (f)	Long form Translation (e)	P(e/f) =freq(e,f)/freq(f)
nt	night	0.1142
nt	note	0.0571
nt	not	0.8285
lv	live	0.125
lv	leave	0.125
lv	love	0.75

Table 7. Bi-gram Probability Distribution Table

Short form Bi-gram (f)	Long form Translation (e)	P(e/f) =freq(e,f)/freq(f)
no i	know i	0.3333
no i	no i	0.6666
n ur	and your	0.4
n ur	in your	0.6
me the	me the	0.3
me the	may the	0.7
u 2	you to	0.5714
u 2	you too	0.4285

Table 8. Analysis of Precision

Sr. No	Number of sample SMS used	Number of Word Tokens (N)	Vocabulary Size (V)	No. of actual output tokens matching with expected tokens (O)	% Precision =O*100/V
1	10	584	76	64	84.210
2	20	1380	139	118	84.892
3	30	2324	204	173	84.803
4	40	3052	246	211	85.772
5	50	3720	288	248	86.111
6	60	4216	322	277	86.024
7	70	4852	362	305	84.254
8	80	5512	410	347	84.634
9	90	6036	457	391	85.557
10	100	6488	461	396	85.900

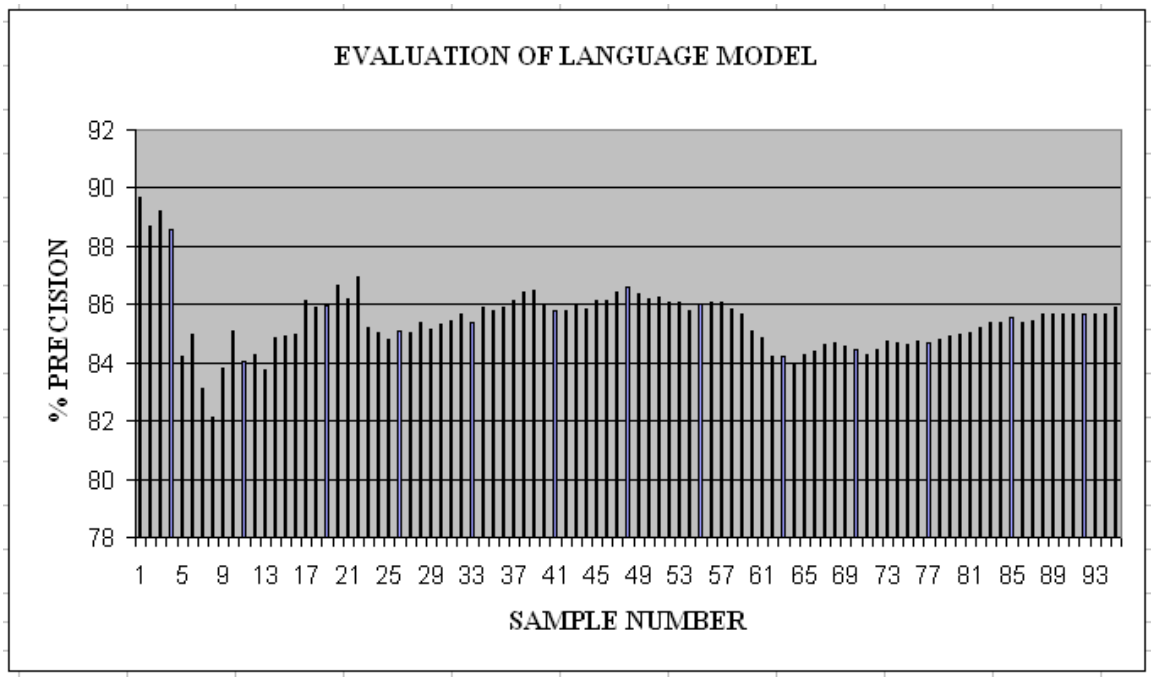

Fig 4: Performance Analysis

4.2 Bi-Gram PDT for the present work

Table 7 shows the Bi-Gram Probability Distribution Table. Here only 8 selected entries out of 2988 entries of Bi-Gram PDT are shown.

4.3 Precision Calculation and Performance Evaluation

Table 8 shows results obtained during the experiment. Here N is number of tokens in samples, V is Vocabulary size and O is number of actual output tokens of decoder matching with expected tokens. Initially 6 sample template messages (not shown in the table) are used to get first result, increasing the number of samples by 1 in each run. The final result is obtained from 100 template messages containing 6488 word tokens. Figure 4 shows the graph obtained from above results while evaluating the language model. It is observed from the graph, change in domain of SMS (sample number 23) and longer length of SMS (sample number 71) degrades the performance of language model. It pretends that training corpus needs to be carefully designed. If the training corpus is too specific to the task or domain, the probabilities may be too narrow and not generalize well to new sentences. If the training corpus is too general, the probabilities may not do sufficient job of reflecting the task or domain. If the different domain sentence is used wile training performance degrades rapidly and again it tries for stability if sentences of same domain are used while training.

Furthermore, if 'test' sentence is part of training corpus or almost similar to any 'train' sentence it produces artificially high probability. Present work has achieved precision of 85.900 % after feeding all 100 template messages, ie. a corpus of 6488 word tokens and a vocabulary of 641 words, to the system and an average performance of 85.4557 %.

5. CONCLUSION

Performance

- There are many Machine Translation models. This work focuses on Bi-gram based statistical LM, which is trained in template messaging domain. SMT systems store different word forms as separate symbols without any relation to each other and word forms or phrases that were not in the training data cannot be translated. This is because of the lack of training data, changes in the human domain where the system is used, differences in morphology age group or mood of person.
- More data tends to yield better language models i.e. predictive accuracy of the language model can be improved by increasing the size of the corpus.
- SMS length is one of the factors affecting the performance. This is because, as the length of SMS grows and the context of text gets clearer, one starts using shorter abbreviations.

Faster Communication

- The developed tool is very useful because it offers substantial savings in human efforts and increased speed of communication.

Contribution of work

- This work provides a facility of combining multiple language words typed with English alphabet.

6. FUTURE SCOPE

- Analyzing the improvement in precision with increased size of Language Model
- Deployment of the software in mobile phones
- Implementation of work using multiplexed PDT

7. REFERENCES

[1] Deana Pennell, Yang Liu, "Toward text message normalization: modeling abbreviation generation", IEEE ICASSP 2011, pp. 5364-5367

[2] Carlos A. Henr´ıquez Q., Adolfo Hern´andez H., "A N-gram based statistical machine translation approach for text normalization on chatspeak style communication", 2009 CAW2.0 2009, April 21, 2009, Madrid, Spain

[3] Waqas Anvar, Xuan Wang, lu Li, Xiao-Long Wang, "A statistical based part of speech tagger for Urdu language", IEEE International Conference on Machine Learing and Cybernetics 2007, 19-22 Aug. 2007, pp. 3418-342.

[4] Srinivas Bangalore, Vanessa Murdock, and Giuseppe Riccardi, "Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system" in 19th International Conference on Computational linguistics, Taipei, Taiwan, 2002, pp. 1–7.

[5] Yong Zhao, Xiaodong He, "Using n-gram based features for machine translation", Proceedings of NAACL HLT 2009: Short Papers, Boulder, Colorado, June 2009, pp. 205–208,

[6] Marcello Federico, Mauro Cettolo, "Efficient handling of n-gram language models for statistical machine translation", Proceedings of the Second Workshop on Statistical Machine Translation, June 2007, Prague, pages 88–95.

[7] Josep M. Crego, Jos´e B. Mari ˜no, "Extending MARIE: an N-gram-based SMT decoder", Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, June 2007, pages 213–216

[8] Zhenyu Lv, Wenju Liu, Zhanlei Yang, "A novel interpolated n-gram language model based on class hierarchy", IEEE International Conference, NLPKE-2009, pp.1-5

[9] Najeeb Abdulmutalib, Norbert Fuhr, "Language models and smoothing methods for collections with large variation in document length", IEEE International Workshop on DEXA-2008, pp. 9-14

[10] Aarthi Reddy, Richard C. Rose, "Integration of statistical models for dictation of document translations in a machine-aided human translation task", IEEE transactions on audio, speech, and language processing, vol. 18, no. 8, November 2010

[11] Evgeny Matusov, "System combination for machine translation of spoken and written language", IEEE transactions on audio, speech, and language processing, vol. 16, no. 7, September 2008

[12] Keisuke Iwami, Yasuhisa Fujii , Kazumasa Yamamoto, Seiichi Nakagawa, "Out-Of-Vocabulary Term Detection By N-Gram Array With Distance From Continuous Syllable Recognition Results", IEEE 2010

[13] Daniel Jurafsky and James H. Martin, "Speech and Language Processing", Pearson Publications, Edition 2011

[14] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer. "The mathematics of statistical machine translation: Parameter estimation", Computational Linguistics, 19(2):263–311, 1993.

[15] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst., "Moses: Open source toolkit for statistical machine translation", In Proceedings of the ACL 2007

[16] J. B. Mari˜no, R. E. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J. A. Fonollosa, and M. R.Costa-juss`a., "N-gram based machine translation", Computational Linguistics, 32(4):527–549,2006.

[17] S. M. Katz. "Estimation of probabilities from sparse data for the language model component of a speech Recognizer". IEEE Trans. Acoust., Speech and Signal Proc., ASSP-35(3), 1987,pp.400–401,

[18] AiTi Aw, Min Zhang, Juan Xian and Jian Su, "A P"A phrase-based statistical model for SMS text normalization," in COLING/ACL, Sydney, Australia, 2006, pp. 33–40.

[19] Catherine Kobus, Francois Yvon, and G´eraldine Damnati, "Normalizing SMS: Are two metaphors better than one?", in 22nd International Conference on Computational Linguistics, Manchester, UK, 2008, pp. 441–448.