Asian Journal of Information Technology 12 (1): 27-32, 2013
ISSN: 1682-3915

Machine Translation Using Multiplexed PDT for Chatting Slang

Rina Damdoo
Department of Computer Science and Engineering,
Ramdeobaba C.O.E.M., Nagpur MS, India

Abstract: A pioneering step in designing Bi-gram based decoder for SMS Lingo SMS Lingo is a language used by young generation for instant messaging for chatting on social networking websites called chatting slang. Such terms often originate with the purpose of saving keystrokes. In last few decades, a significant increment in both the computational power and storage capacity of computers have made possible for Statistical Machine Translation (SMT) to become a concrete and realistic tool. But still it demands for larger storage capacity. My past research employs Bi-gram Back-off Language Model (LM) with a SMT decoder through which a sentence written with short forms in an SMS is translated into long form sentence using non-multiplexed Probability Distribution Tables (PDT). Here, in this study the same is proposed using multiplexed PDT (a single PDT) for Uni-gram and Bi-gram, so smaller memory requirements. Use of N-gram LM in chatting slang with multiplexed PDT is the objective behind this work. As this application is meant for small devices like mobile phones, researchers can prove this approach a memory saver.

Key words: Statistical Machine Translation (SMT), Bi-gram, multiplexed Probability Distribution Table (PDT), parallel alligned corpus, Bi-gram matrix

INTRODUCTION

While messaging SMS one tries to type maximum information in single SMS. This practice has evolved a new language SMS Lingo. Internet users have popularized Internet slang or chatting slang or netspeak or chatspeak a type of slang that many people use for texting on social networking websites to speed up the communication. Very few people now a days write you for you than u for you. Such terms often originate with the purpose of saving keystrokes. Secondly young generation does not pay attention to grammar like instead writting I am waiting they write am waiting or I waiting or me waiting.

Thirdly the consequence of using his casual language is word based translation model does fail if a person uses same abbreviation for more than one word. Because from the data corpus collected it is observed one writes same abbreviation "wh" some times for "what" some times for "where" some times for "why" sometimes for who so to get the context clearer he earlier and/or later words also must be considered. In short, a context analysis evaluation should be made to choose the right definition (Pennell and Liu, 2011; Henriquez and Hernandez, 2009; Anwar *et al*., 2007; Bangalore *et al*., 2002). Table 1 gives some sample abbreviations with their expanded definitions. Figure 1 shows chatting slang in an example session of two persons on social networking websites. Both user A and B are typing short text but the end user is able to see the long form text which increases the readability. Like this text normalization (Pennell and Liu, 2011) patent and reference searches and various information retrieval systems kids' self learning can be main applications of this kind.

Table 1: Sample abbreviations with their multiple expanded definitions

Abbreviations	Expanded definitions
lt	Let, Late
the	The, There, Their
n	In, And
me	Me, May
wer	Were, Wear
dr	Dear, Deer, Doctor

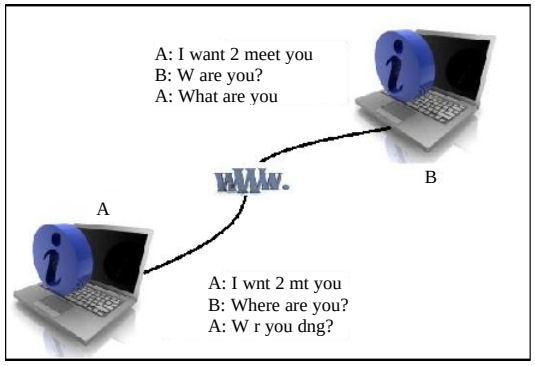

Fig. 1: Example session of two persons on internet

The earlier research (Damdoo and Shrawankar, 2012a, b) employs Bi-gram LM with Back-off SMT

decoder for template messaging through which a sentence written with short forms Sentence (S) in an SMS is translated in to long form sentence (L) using non-multiplexed Probability Distribution Tables (PDT) Bi-gram PDT and Uni-gram PDT. The software performs following steps:

- Data corpus collection
- Preprocessing the corpus
- Training the LM
  - Generating Uni-gram and Bi-gram PDTs
- Testing the LM
  - Using Uni-gram and Bi-gram PDTs
  - Using back-off decoder to expand short SMS to long SMS
- Evaluating LM with performance and correctness measures
  - Precision recall and F-factor

While working on this project it was experienced PDT designing and generation is the most important phase in the project. Because as this application is meant for small devices like mobile phones memory usage is of most concern. In this research work using multiplexed PDT (Single PDT) for Uni-gram and Bi-gram is presented.

N-GRAM BASED SMT SYSTEM

Among the different machine translation approaches, the statistical N-Gram-Based System (Henriquez and Hernandez, 2009; Zhao and He, 2009; Federico and Cettolo, 2007; Reddy and Rose, 2010; Marino *et al*., 2006; Katz, 1987) has proved to be comparable with the state of the artphrase based systems (Koehn *et al*., 2007).

The SMT probabilities a the sentence level are approximated from word-based translation models that are trained by using bilingual corpora (Iwami *et al*., 2011) and in an N-gram LM N-1 words are used to predict the next Nth word. In Bi-gram LM (N = 2) only previous word is used to predict the current word. SMT has two major components (Anwar *et al*., 2007; Crego and Marino, 2007; Iwami *et al*., 2011; Jurafsky and Martin, 2011):

- A probability distribution table
- A Language Model decoder

A PDT captures all the possible ranslations of each source phrase. These translations are also phrases. Phrase tables are created heuristically using the Word-Based Models (Brown *et al*., 1993). The probability of target phrase if f is source and e is target language is given as:

$$P\left(\frac{L}{S}\right) = P\left(\frac{e}{f}\right) = \frac{\text{Count}(e,f)}{\text{Count}(f)}$$

$$P\left(\frac{\text{too}}{2}\right) = \frac{\text{Count}(\text{too},2)}{\text{Count}(2)} = \frac{2}{10} = 0.20$$

This means Uni-gram 2 is present ten times in the collected corpus out of which only twice it represents too. In an N-gram model the probability of a word is approximated given all the earlier n words by the conditional probability of all the preceding words $P(w_n \backslash w_1^{n-1})$. The Bi-gram model approximates the probability of a word given the earlier word $P(w_n \backslash w_{n-1})$:

$$P\left(\frac{w_n}{w_{n-1}}\right) = \frac{\text{Count}(w_{n-1}w_n)}{\text{Count}(w_{n-1}w)}$$

Researchers can simplify his equation since sum of all Bi-gram counts that start with a given word w_{n-1} must be equal to the Uni-gram count for that word w_{n-1}.

$$P\left(\frac{w_n}{w_{n-1}}\right) = \frac{\text{Count}(w_{n-1}w_n)}{\text{Count}(w_{n-1})}$$

The unsmoothed (there are no unknown words). Maximum likelihood estimate of uni-gram probability can be computed by dividing the count of the word by the total number of word takens N (Federico and Cettolo, 2007; Jurafsky and Martin, 2011):

$$P(w) = \frac{\text{Count}(w)}{\sum_i \text{Count}(w_i)}$$

$$P(w) = \frac{\text{Count}(w)}{N}$$

As probabilities are all <1 the product of many probabilities (Probability chain rule) gets smaller the more probabilities one multiply. This causes a practical problem of numerical underflow (Matusov *et al*., 2008; Jurafsky and Martin, 2011). In this case it is customary to do the computation in log space take log of each probability (the logprob) in computation. But the PDT still contains the probabilities or word counts.

PHRASE TABLE OR PROBABILITY DISTRIBUTION TABLE WITHOUT MULTIPLEXING

Due to the lack of enough amount of raining corpus word probability distribution (Abdulmutalib and Fuhr, 2008; Jurafsky and Martin, 2011) is misrepresented.

Table 2: Uni-gram PDT for the corpus in Table 3

Uni-gram (w)	Uni-gram probability count (w)/N	Uni-gram (w)	Uni-gram probability count(w)/N
i	10/120 = 0.083	where	2/120 = 0.016
want	3/120 = 0.025	w	6/120 = 0.050
wnt	2/120 = 0.016	wh	2/120 = 0.016
wan	5/120 = 0.041	whe	2/120 = 0.016
to	6/120 = 0.050	are	9/120 = 0.075
2	4/120 = 0.330	r	11/120 = 0.091
meet	2/120 = 0.016	what	3/120 = 0.025
mt	4/120 = 0.033	wht	5/120 = 0.041
met	4/120 = 0.033	doing	4/120 = 0.033
you	12/120 = 0.100	dng	2/120 = 0.016
u	18/120 = 0.150	dong	4/120 = 0.030

Table 3: Source and test data

Long form language (L)/ target language (e)	Short form language (S)/source language (f)
I want to meet u.	I want to meet you. Where are you?
Wh are u?	What are you doing?
What r u doing?	I wnt 2 mt u. W are u? W r u dng?
	I wan 2 mt u. Whe r u? Wht r u doing?
	I wan 2 met u. Where are u? Wht r u dong?
	I want 2 meet u. W r u? What are you dng?
	I wnt to mt you W r u? W are you dong?
	I wan 2 mt you. Whe are u? Wht are you doing?
	I wan to meet you. Wh r you? What r you doing?
	I wan to met you. Where r u? Wht are you dong?
	I want 2 met you. W r u? What are you dong?

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. The result is a separation of a Bi-gram into two Uni-grams.

Uni-gram PDT: Table 2 shows uni-gram PDT for the corpus in Table 3. For this corpus N = 120. From this PDT, it is observed Uni-gram u occurs 18 times in the collected corpus in Table 3 hence has the highest probability of 0.15.

Bi-gram PDT: Table 4 shows Bi-gram PDT for the corpus in Table 3. From this PDT it is observed Bi-gram "r u" occurs 9 times and 'r you' occurs 2 times which predicts that in this kind of short form language chances of a person to write "r u" is more than to write "r you" hence, "r u" has the higher probability of 0.81 over "r you" with probability 0.18.

PROPOSED WORK

One can create a single matrix of Bi-grams instead of two separate PDT tables for Uni-gram and Bi-gram a multiplexed PDT (Jurafsky and Martin, 2011). Table 5 shows a multiplexed PDT for the corpus in Table 3. Unlike un-multiplexed PDT this PDT contains the Bi-gram counts. The reason behind this is provision to calculate probability of a Uni-gram from the same PDT. This PDT (matrix) is of size (V+1)×(V+1) where V is the total number of word types in the language the

Table 4: Bi-gram PDT for the corpus in Table 3

Bi-gram (w_1w_2)	Bi-gram probability count (w_1w_2)/count (w_1)	Bi-gram (w_1w_2)	Bi-gram probability count (w_1w_2)/count (w_1)
i wnt	2/10 = 0.20	met u	2/4 = 0.50
wnt 2	1/2 = 0.50	where are	1/2 = 0.50
2 mt	3/4 = 0.75	u dong	1/18 = 0.05
mt u	2/4 = 0.50	2 meet	1/4 = 0.25
w are	2/6 = 0.33	what are	2/3 = 0.66
are u	4/9 = 0.44	are you	5/9 = 0.55
w r	4/6 = 0.66	you dng	1/12 = 0.08
r u	9/11 = 0.81	wnt to	1/2 = 0.50
u dng	1/18 = 0.05	to mt	1/6 = 0.16
i wan	5/10 = 0.50	mt you	2/4 = 0.50
wan 2	3/5 = 0.60	you dong	3/12 = 0.25
whe r	1/2 = 0.50	whe are	1/2 = 0.50
wht r	2/5 = 0.40	wht are	2/5 = 0.40
u doing	2/18 = 0.11	you doing	2/12 = 0.16
i want	3/10 = 0.30	wan to	2/5 = 0.40
want to	2/3 = 0.66	to met	2/6 = 0.66
to meet	1/6 = 0.16	met you	3/4 = 0.75
meet u	1/2 = 0.50	wh r	1/2 = 0.50
wh are	1/2 = 0.50	r you	2/11 = 0.18
what r	2/3 = 0.66	where r	1/2 = 0.50
2 met	2/4 = 0.50	want 2	1/3 = 0.66

vocabulary size. <S> is a special Uni-gram used in between the sentences (as start of sentence or end of sentence). This special Uni-gram plays an important role to find the context of a sentence. To see the Bi-gram count corresponding row for the first word and the count in the corresponding column for the second word in Bi-gram is seen. From the PDT probability of Bi-gram 'r u' is calculated as follows:

- Uni-gram count of r is found by adding all the entries of r row which is 11:

$$P(r) = \frac{\text{Count (r)}}{120} = \frac{11}{120} = 0.091$$

- In the row of Uni-gram 'r' and column of Uni-gram 'u' count is 9:

$$P(r) = \frac{\text{Count (r u)}}{\text{Count (r)}} = \frac{9}{11} = 0.81$$

Majority of the values are zero in this matrix (sparse matrix) as the corpus considered is limited. As the size of the corpus grows one gets more combinations of word tokens as Bi-grams (out of the scope of this study).

EXPERIMENTAL SETUP

The project is divided in to two phases:

- Multiplexed PDT generation
- Implementation of Back-off decoder using multiplexed PDT

Table 5: Multiplexed PDT for the corpus in Table 3

Multiplexed PDT	I	want	wnt	wan	to	2	meet	mt	met	you	u	where	w	wh	whe	are	r	what	wht	doing	dng	dong	<S>
I	0	4	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
want	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wnt	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wan	0	0	0	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
to	0	0	0	0	0	0	1	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
meet	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
mt	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0
met	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0
you	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	3	6
u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	14
where	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	0	0	0	0	0	0
wh	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
whe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
are	0	0	0	0	0	0	0	0	0	5	4	0	0	0	0	0	0	0	0	0	0	0	0
r	0	0	0	0	0	0	0	0	0	2	9	0	0	0	0	0	0	0	0	0	0	0	0
what	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0
wht	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	0	0	0	0	0	0
doing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
dng	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
dong	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
<S>	10	0	0	0	0	0	0	0	0	0	0	2	6	2	2	0	0	4	4	0	0	0	0

Table 6: Multiplexed PDT for the corpus in Table 3 to contain additional information about Bi-gram

Multiplexed PDT	I	want	wnt	wan	to	2	meet	mt	met	you	u	where	w	wh	whe	are	r	what	wht	doing	dng	dong	<S>
I	0	X	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
want	0	0	0	0	X	1	0	0	0	X	0	0	0	0	0	0	0	0	0	0	0	0	0
wnt	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
wan	0	0	0	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
to	0	0	0	0	0	0	1	3	2	X	0	X	0	0	0	0	0	X	0	0	0	0	0
2	0	0	0	0	0	0	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
meet	0	0	0	0	X	0	0	0	0	X	1	X	0	0	0	0	0	X	0	0	0	0	X
mt	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0
met	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0
you	0	0	0	0	X	0	X	0	0	0	0	0	0	0	0	X	0	X	0	2	1	3	X
u	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	1	14
where	X	0	0	0	X	0	0	0	0	X	0	0	0	0	0	X	1	0	0	0	0	0	0
w	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	4	0	0	0	0	0	0
wh	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
whe	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
are	0	0	0	0	X	0	0	0	0	X	4	X	0	0	0	0	0	X	0	0	0	0	0
r	0	0	0	0	0	0	0	0	0	2	9	0	0	0	0	0	0	0	0	0	0	0	0
what	X	0	0	0	X	0	0	0	0	X	0	0	0	0	0	X	1	0	0	0	0	0	0
wht	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	3	0	0	0	0	0	0
doing	0	0	0	0	X	0	X	0	0	0	0	X	0	0	0	X	0	X	0	0	0	0	X
dng	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
dong	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
<S>	X	0	0	0	0	0	X	0	0	X	0	X	6	2	2	0	0	X	4	0	0	0	0

In development and testing process in first phase data for the project work is collected from 10 persons which is each of 1500 words. Table 3 shows a piece of the data collected. This data is used to train the LM to get a multiplexed PDT. Before providing word-aligned parallel corpus to first phase it is preprocessed by removing extra punctuation marks extra spaces and representing begin and end of statement by <S>.

This is done using regular expression meta characters in JAVA. This is useful in context checking. Table 6 shows multiplexed PDT with some additional information about Bi-gram required in soft ware in which along with the information of probability we need to know the long form for the Bi-gram. This information is kept in the same matrix with a link field a common link for all the source short form Bi-grams having the same arge long form ranslation. Some do-not are (X) entries in this table are also used. These entries are used to save the time of back-off decoder. While looking for a Bi-gram as soon as the decoder finds X it copies the input phrase to the out put string without going for the further calculation of probability. Otherwise if the decoder is unsuccessful to

find non-zero entry for the Bi-gram it breaks it in to two Uni-grams. These Uni-grams are then separately handled by the decoder.

There is additional link field for targe Uni-gram long form translation. If the decoder is unable to find the Uni-gram in the PDT it copies the input word as it is to the output string.

EXPERIMENTAL RESULTS

This software produces correct translations for the seen words and unseen words are output without any alteration. Also for some Bi-grams like "w r" the results depends on the indexing of the PDT. For example "w r" always produced where are as word token "w" in the corpus appeared first time for long word where. This limitation can be overcome by making more than one entries for word token "w" one when I appear in place of where and another when it appears in place of what. Word combinations like lol for lots of love can not be expanded as the work is limited to word aligned parallel corpus. Finally implementation point view creation and handling of multiplexed PDT is more complex as compared to separate PDTs in machine translation application.

CONCLUSION

This researh focuses on multiplexed PDT Bi-gram based statistical LM which is trained in chatting slang language domain. SMT Systems store differen word forms as separate symbols without any relation to each other and word forms or phrases that were no in the raining data cannot be translated. As his application is mean for small devices like mobile phones researchers can prove this approach a memory saver. In future the research can be done on performance improvement by increasing the size of the corpus and the language model using multiplexed PDT. Patent and reference searches and various information retrieval systems communication on social networking websites are the main applications of the work.

REFERENCES

Abdulmutalib, N. and N. Fuhr, 2008. Language models and smoothing methods for collections with large variation in document length. Proceedings of the 19th International Workshop on Database and Expert Systems Application, September 1-5, 2008, Turin, pp: 9-14.

Anwar, W., X. Wang, L. Li and X.L. Wang, 2007. A statistical based part of speech tagger for Urdu language. Proceedings of the International Conference on Machine Learning and Cybernetics, August 19-22, 2007, Hong Kong, China, pp: 3418-3424.

Bangalore, S., V. Murdock and G. Riccardi, 2002. Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. Proceedings of the 19th International Conference on Computational linguistics, August 26-30, 2002, Taipei, Taiwan, pp: 1-7.

Brown, P.F., V.J.D. Pietra, S.A.D. Pietra and R.L. Mercer, 1993. The mathematics of statistical machine translation: Parameter estimation. Compuaional Linguisics, 19: 263-311.

Crego, J.M. and J.B. Marino, 2007. Extending MARIE: An N-gram-based SMT decoder. Proceedings of the Annual Meeting of the Association of Computational Linguistics System Demonstration, June 2007, Prague, pp: 213-216.

Damdoo, R. and U. Shrawankar, 2012a. Probabilistic N-gram language model for SMS Lingo. Proceedings of the International Conference on Recent Advances in Computing and Software Systems, April 25-27, 2012, Chennai, India, pp: 114-118.

Damdoo, R. and U. Shrawankar, 2012b. Probabilistic language model for template messaging based on Bi-gram. Proceedings of the International Conference on Advances in Engineering, Science and Management, March 30-31, 2012, Nagapattinam, Tamil Nadu, pp: 196-201.

Federico, M. and M. Cettolo, 2007. Efficient handling of N-gram language models for statistical machine translation. Proceedings of the 2nd Workshop on Statistical Machine Translation, June 2007, Prague, pp: 88-95.

Henriquez, Q.C.A. and H.A. Hernandez, 2009. A Ngram-based statistical machine translation approach for text normalization on chat-speak style communications. CAW2.0 2009, Madrid, Spain. http://www2009.eprints.org/255/3/Henriquez_Hernandez_CAW2009.pdf.

Iwami, K., Y. Fujii, K. Yamamoto and S. Nakagawa, 2011. Efficient out-of-vocabulary term detection by n-gram array indices with distance from a syllable lattice. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 22-27, 2011, Prague, pp: 5664-5667.

Jurafsky, D. and J.H. Martin, 2011. Speech and Language Processing. 2nd Edn., Prentice Hall, New Jersey, USA.

Katz, S., 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. IEEE Trans. Acoustics Speech Signal Process., 35: 400-401.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch and M. Federico *et al*., 2007. Moses: Open source toolkit for statistical machine translation. Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, June 2007, Prague, pp: 177-180.

Marino, J.B., R.E. Banchs, J.M. Crego. A. de Gispert, P. Lambert, J.A.R. Fonollosa and M.R. Costa-jussa, 2006. N-gram based machine translation. Comput. Ling., 32: 527-549.

Matusov, E., G. Leusch, R.E. Banchs, N. Bertoldi and D. Dechelotte *et al*., 2008. System combination for machine translation of spoken and written language. IEEE Trans. Audio Speech Lang. Process., 16: 1222-1237.

Pennell, D. and Y. Liu, 2011. Toward text message normalization: Modeling abbreviation generation. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 22-27, 2011, Prague, Czech Republic, pp: 5364-5367.

Reddy, A. and R.C. Rose, 2010. Integration of statistical models for dictation of document translations in a machine-aided human translation task. IEEE Trans. Audio Speech Lang. Process., 18: 2015-2027.

Zhao, Y. and X. He, 2009. Using N-gram based features for machine translation system combination. Proceedings of the Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, June 2009, Boulder Colorado, pp: 205-208.