International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 5 Issue: 5
ISSN: 2454-4248
128– 132

Analysis of Topic Modeling on Phrase-Based SMT system for English-Hindi Translation

Himanshu Sharma[1], Harvir Singh[2]
[1] Research Scholar, Jaipur National University, Jaipur, Rajasthan, India
[2] Director, School of Computer and Systems Sciences, Jaipur National University, Jaipur, Rajasthan, India
Email- himsharma14@gmail.com, Dr.harvir@gmail.com

Abstract :- After availability of cheaper large memory and high performance processors, Statistical Machine Translation (SMT) methods have drawn attention of researchers NLP. Phrase-based SMT has shown better results than word-based SMT. To improve performance of machine translation system further, different systems have been developed which use phrase-based SMT as a baseline system. Domain adaptation is one of most popular example of such systems. In this paper also phrase-based SMT system is used as baseline to apply topic model for English-Hindi translation. This baseline system is also used for result comparison with topic model system. Both systems are trained using MERT. The analysis shows improvement in results obtained by using topic modeling system.

Keywords – *Phrase-based SMT, Topic modeling, LDA.*

*****

1. INTRODUCTION

Efforts are increasingly made by researchers to use Machine Translation in NLP across the globe especially since last two decades. Availability large memory and high performance processors has also motivated the researchers to focus their efforts in the area. Similar efforts are also being made for Indian languages too [1], [2], [3] & [4]. Translation between English and Indian languages has also attracted researchers' attention. Different approaches like rule based [5], interlingua based [6] and statistical methods [7] have also been explored for English-Hindi translation.

A large number research have been exploring use of domain adaptation to improve results of SMT. In this paper, we will explore effect of using topic modeling based approach for English-Hindi machine translation. As a baseline system, we have used phrase-based SMT that will also be used for performance comparison with our topic modeling based system. The topic modeling system is then incorporated to the baseline SMT. The idea of using topic modeling to the baseline SMT is to give higher probabilities to target phrase according to the source phrase's topic.

By applying both baseline SMT and topic model systems to the discussed in the next section we observed encouraging performance improvements.

The section 2 discusses different English-Hindi parallel corpus. Section 3 will describe the phrase-based SMT system used as baseline. In section 4, we will describe topic model to be used whereas section 5 describes how this topic model is integrated with the baseline system.

2. CORPORA

The parallel corpus can be characterized in three ways, namely number of languages, direction of translation and level of alignment. Number of parallel data extraction techniques focus on the second-order class of the layout. The work done by Baker et al. [8] EMILLE portrays corpus with regard to collection of parallel of data. It highlights the challenges of collecting PDFs and images. It is one of EMILLE/CIIL which are from a few early corpuses known and developed by Lancaster University in collaboration of Central Institute of Indian Languages, India through the EMILLE project. It consists of texts from English and Indian languages – Hindi, Bengali and 3 other. EMILLE /CIIL corpus covers three domains namely health, legal and education. As a result of language contest on SMT 2002, another corpus set, known as DARPA-TIDES, was also developed for use in English<-> Hindi translation.

Department of Information and Technology of India initiated two different projects, English to Indian languages MT (EILMT) and Indian Languages Corpora Initiative (ILCI). These projects focus on development of parallel corpus for English and different languages. Both of these projects collect resources relating to two domains: health and tourism. EILMT also contains domain-specific term translations and multi-word expressions for both the domains. Whereas linguistic annotators created POS (part-of-speech) tags are included in ILCI.

 ISSN: 2454-4248
128– 132

Chaudhury et al. [9] contributed parallel text in many languages other than English and Hindi which is included in GyanNidhi corpus. Alignment in the work was done using heuristic approach based on MT. Different problems encountered and technique to collect English-Hindi parallel corpus were discussed by Bojar et al. [10]. Singh and Bandyopadhyay (2010) explain how to edit PDF documents for UTF-8 design, while information is accumulated for the English-Manipuri pair, mainly for newspapers.

Crowd-sourcing is also used to create parallel corpus in Indian languages [11]. This was built-up by using translators to translate from Indian languages to Hindi. It provides multiple alternate translations for one sentence in Indian language.

An attempt was also made by mixing parts of TIDES, Tourism-EILMT and EMILLE-ACL05 corpora [12]. This resulted in a large corpus in English and Hindi. Earlier version of the corpus had issues relating to quality of source datasets.

Table 1 summarizes statistics of various corpus available.

Corpus	No. of sentences	No. of En token	No. of Hi token
EMILLE-ACL05	3,556 57	57,118	70,932
TIDES-ICON08	52,000	12,43,815	13,38,994
Tourism-EILMT	15,198	3,83,992	3,65,163
Health-EILMT	7,484	1,37,396	1,69,039
Tourism-ILCI	25,000	4,25,646	4,23,711
Health-ILCI	25,000	4,22,436	4,40,764
NCERT	9,340	1,73,129	1,98,264
Total	137,578	-	-

Table 1 Statistics of the different datasets

3. PHRASE BASED BASELINE SYSTEM

The problem of finding the best translation e_{best} of a source statement f can be modeled in terms of maximum posterior probability [13]. The model can be represented using eq. (1).

$$e_{best} = argmax_e P(e|f)$$
$$= argmax_e P(f|e)P_{lm}(e) \quad (1)$$

Here $P(e|f)$ is a translation model and P_{lm} is the language model.

The baseline phrase-based SMT used in the system follows the model by Koehn et al. [14] to adapt following six features-

1. Two phrase translation probabilities (one for each direction).
2. Two word translation probabilities (one for each direction).
3. Target language model.
4. Distance-based model.
5. Phrase penalty for target language
6. One word penalty for target language

The above features are then log-linearly interpolated [15] using the following formula-

$$e_{best} = argmax \{\sum_{m=1}^{M} \lambda_m h_m(e,f)\} \quad (2)$$

Where λ_m is the weight optimized using a discriminative training method on development data and $h_m(e,f)$ is a feature function.

The phrase-based SMT may suggest more than one target phrases for a single source phrase with different probabilities. To select the best translation, the system uses topic modeling approach that assigns probabilities to the target phrases based on biased weight to the specific sub-model according to the specific domain of the source phrase and then combines it with a general model.

4. IMPLEMENTING TOPIC MODELING TO THE BASELINE SYSTEM

Though there are several various methods available for topic modeling that can be used in NLP, but LDA [16] is the most popular topic model used.

LDA provides distribution of a topic over the words, i.e. the word-topic distribution $p(word_j | topic_i)$ during training by using clustering. In addition, LDA has inference ability like a classifier. Using the word-topic distributions, provided by LDA, for various topics in both source and target languages, topic distribution $p(topic_i|doc_{new})$ for a text is derived. Thus topic of a source text can be detected using LDA. Same

 ISSN: 2454-4248
128– 132

way, topic distribution for a target phrase can be detected by averaging the word-topic distributions over all the words in the target phrase.

Having topics of source and target corpus using LDA, topic model is applied to baseline phrase-based SMT using following steps.

I. Mapping Source and target topic models –

Source and target topic models are then mapped using following steps-

1. Perform word alignment on special bilingual corpus in two directions using GIZA++.
2. For each topic in both languages, top-n (n=200) word-topic distributions are chosen.
3. Then, mapping words between different combinations of topics in both languages are counted and sum their distribution values to determine the mapping.

II. Apply topic model

Topic modeling obtained in step 3 then incorporated in the following manner-

1. The translation model is trained using MERT [17] to get a phrase table covering all topics. Parallel corpus is used for training translation model.
2. Topic of the source input text, T_s, is found out. Thereafter the corresponding target language topic, Tg,is found out by looking up source-to-target topic mapping table.
3. To assign higher probability to target phrase related to a specific topic following method is used-

a) Let PS_t is target phrase in Topic Tg and PS_t is made up of $\{W_1, W_2, \ldots, W_N\}$

b) Topic relevance for PS_t is calculated by-

$$Rel(PS_t, T_g) = \frac{\left(\sum_{j=1}^{N} p(W_j, T_g)\right)}{N} * P(T_g) \quad (3)$$

c) Value of $P(T_g)$ can be calculated in following steps-

i. Find out source topic of the source input text (i.e. T_s for new document doc_{new})

ii. Find out target topic T_g corresponding to T_s.

iii. Calculate $P(T_g)$ using-

$$P(T_g) = \max(p(t_i|doc_{new})) \quad (4)$$

where $i = 1,2,\ldots, H$ *is number of topics*

5. RESULTS

Though experiment was performed for 10 topics in both languages, the below table shows word-topic distribution for 4 topics of Hindi corpus only. The table shows word-topic distribution in target language (Hindi) with 5 top priority words for each topic-

Topic 1		Topic 2		Topic 3		Topic 4	
Word	P(w|t)	Word	P(w|t)	Word	P(w|t)	Word	P(w|t)
काम	0.04077	नस्ल	0.03998	नस्ल	0.02506	काम	0.02877
नस्ल	0.02944	रूप	0.02616	साफ	0.0208	मुर्गीपालन	0.01522
भारतीयों	0.01115	साफ	0.01133	मुख्य	0.01399	साल	0.0142
खरीदने	0.00566	काला	0.00668	केरल	0.00961	लगता	0.00984
मुख्य	0.00542	घास	0.00585	काला	0.00806	वर्षों	0.00821

Table 2 Word-topic distribution in target language (Hindi)

System	LM(e)	Pphr(e|f)	Pw(e|f)	Pphr(f|e)	Pw(f|e)	PP(f)	WP(e)	Rel(e)
Baseline	0.5861	0.0951	0.0894	0.0996	0.1894	0.1703	-0.3015	
Topic Model	0.325	0.0742	0.1097	0.0627	0.1057	0.116	-0.2614	0.318

Table 3 Weights of various features obtained by training using MERT

Both the systems, baseline and topic model, were trained on general corpus. MERT was used to fine-tune weights of various features for both baseline and topic modeling systems and relevance feature (Rel(e)) used in topic modeling system. As it can be observed from the Table 3 that weights of Rel(e) and LM are approximately equal.

 ISSN: 2454-4248
128– 132

Language model was implemented for upto 4-grams and the results were evaluated using BLEU and NIST scores. The tests were carried out on Language model was implemented and the results were evaluated using BLEU and NIST scores. The tests were carried out on both general corpuses. Table 4 shows the details of the results obtained.

It can be observed from the results in Table 4 that both BLEU and NIST scores are improved when Topic modeling system is applied to the same general corpus which is used for baseline system.

Corpus	System	1-gram	2-gram	3-gram	4-gram	BLEU	NIST
General Corpus	**Baseline**	57.69	28.845	14.4225	7.21125	18.76	5.869
	Topic Model	58.84	29.42	14.71	7.355	19.23	5.996

Table 4 BLEU and NIST scores of Baseline and Topic Modeling systems

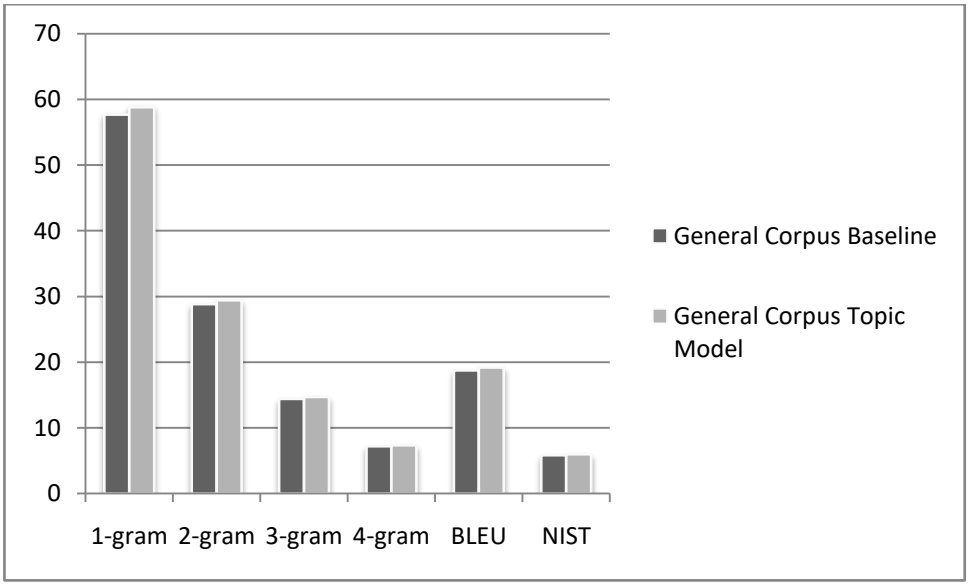

Figure 1 BLUE and NIST scores applied on General Corpus

6. Conclusion and Future Work

In this paper we incorporated Topic model to phrase-based statistical machine translation and applied to different general corpuses. We found that despite the fact the corpses were not specific to any specific structure the translation quality of English-Hindi translation is improved when Topic modeling is integrated to phrase-based SMT.

In future, we will explore effect of using topic modeling system for the same set of languages after general corpus with special in-domain corpus.

References

[1] Ramanathan, A., Choudhary, H., Ghosh, A., and Bhattacharyya, P. (2009). Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint in proceedings on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 800–808, Suntec, Singapore, August. Association for Computational Linguistics.

[2] Venkatapathy, S. and Bangalore, S. (2009). Discriminative Machine Translation Using Global Lexical Selection. *ACM Transactions on Asian Language Information Processing*, 8(2).

[3] Arafat, A., Kolachina, P., Kolachina, S., Sharma, D. M., and Sangal, R. (2010). Coupling Statistical Machine Translation with Rule-based Transfer and Generation. In *Proceedings of the 9th Conferenceof the Association for Machine Translation in the Americas (AMTA)*.

[4] Venkatapathy, S., Sangal, R., Joshi, A., and Gali, K. (2010). A Discriminative Approach for Dependency Based Statistical Machine Translation. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 66–74, Beijing, China, August. Coling 2010 Organizing Committee.

[5] Sinha, RMK and Sivaraman, K and Agrawal, A and Jain, R and Srivastava, R and Jain, A. 1995. ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages IEEE International Conference on Systems, Man and Cybernetics

[6] Dave, Shachi and Parikh, Jignashu and Bhattacharyya, Pushpak. 2001. Interlingua-based English–Hindi Machine Translation and Language Divergence Journal Machine Translation

 ISSN: 2454-4248
128– 132

[7] Ananthakrishnan Ramanathan, Pushpak Bhattacharyya, Jayprasad Hegde, Ritesh M. Shah, and M. Sasikumar. 2008. Simple syntactic and morphological processing can help English-Hindi statistical machine translation. In International Joint Conference on NLP.

[8] P. Baker, A. Hardie, T. McEnery, R. Xiao, K. Bontcheva, H. Cunningham, R. Gaizauskas, O. Hamza, D. Maynard, V. Tablan, et al., "Corpus linguistics and south asian languages: Corpus creation and tool development," *Literary and Linguistic Computing*, vol. 19, no. 4, pp. 509–524, 2004.

[9] S. Chaudhury, D. M. Sharma, and A. P. Kulkarni, "Enhancing effectiveness of sentence alignment in parallel corpora: Using mt heuristics," *ICON, vol. 29*, 2008.

[10] Bojar, O., Stranák, P., and Zeman, D. (2010). Data Issues in English-to-Hindi Machine Translation. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

[11] Post, M., Callison-Burch, C., and Osborne, M. (2012). Constructing Parallel Corpora for Six Indian Languages via Crowdsourcing. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 401–409, Montréal, Canada, June. Association for Computational Linguistics.

[12] Bojar, O., Diatka, V., Rychlý, P., Stranák, P., Tamchyna, A., and Zeman, D. (2014). Hindi-English and Hindi-only Corpus for Machine Translation. In *Proceedings of the Ninth International Language Resources and Evaluation Conference (LREC'14)*, Reykjavik, Iceland, May. ELRA, European Language Resources Association. in prep.

[13] PF Brown, SA Della Pietra, VJ Della Pietra, RL Mercer. 1992. The Mathematics of Statistical Machine Translation: Parameter Estimation. Computational Linguistics. 19(2):263-309.

[14] Philipp Koehn, Franz . Josef Och, and Daniel Marcu.2003. Statistical Phrase-Based Translation .Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, pages 48-54.

[15] Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In Proc. of ACL00, pages 440–447.

[16] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning Research, pages 993–1022.

[17] Franz Josef Och. Minimum Error Rate Training in Statistical Machine Translation. 2003. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 160–167.