

How Sentiment Analysis Can Help Machine Translation

Santanu Pal[‡], Braja Gopal Patra[‡], Dipankar Das[‡], Sudip Kumar Naskar[‡],
Sivaji Bandyopadhyay[†] and Josef van Genabith^{*}

^{*}Universität des Saarlandes, Saarbrücken, Germany

[†]Jadavpur University, Kolkata, India

^{*}{santanu.pal, josef.vangenabith}@uni-saarland.de

[†]{brajagopal.cse, dipankar.dipnil2005}@gmail.com

[‡]{sudip.naskar, sbandyopadhyay}@cse.jdvu.ac.in

Abstract

State-of-the-art Machine Translation (MT) does not perform well while translating sentiment components from source to target language. The components such as the sentiment holders, sentiment expressions and their corresponding objects and relations are not maintained during translation. In this paper, we described, how sentiment analysis can improve the translation quality by incorporating the roles of such components. We also demonstrated how a simple baseline phrase-based statistical MT (PB-SMT) system based on the sentiment components can achieve 33.88% relative improvement in BLEU for the under-resourced language pair English-Bengali.

1 Introduction

The Statistical machine translation (SMT) systems are considered as one of the most popular approaches to machine translation (MT). However, SMT can suffer from grammatically incorrect output with erroneous syntactic and semantic structure for the language pair on which it is being applied. It is observed that the grammatical errors not only weaken the fluency, but in some cases it may even completely change the meaning of a sentence. In morphologically rich languages, grammatical accuracy is of particular importance, as the interpretation of syntactic relations depends heavily on the morphological agreement within sentences. Morphological errors create serious problems in context of translating the sentiment related components from source to target language. In this paper, we handle these errors by focusing on the roles of sentiment-holder, sentiment expression,

corresponding objects and their relations with each other at the clause level.

A common error that occurs during translation using SMT is the relations among the holders, associated sentiment expressions and their corresponding objects in a sentence (in case of complex and compound sentences) may interchange. In the following example, the position of the sentiment expression has been changed in target language while translated. Similar instances are found if any interchange occurs in case of other sentiment components such as holder or object.

Example 1: Source: In 1905, <holder>Calcutta</holder> <expression_1>protested </expression_1> <object_1>the partition of Bengal</object_1> and <expression_2>boycotted </expression_2> <object_2>all the British Goods </object_2>.

Target: 1905 sale, <holder>Calcutta </holder> <object_1>bongo vongo-r</object_1> <expression_2>boycot korechilo </expression_2> <object_2>ssmosto british samogri </object_2> <expression_1> protibad janiyechilo </expression_1>.

Thus, the entire semantics of the sentence has been changed even the sentence is considered as grammatically correct. Another major challenge is to develop a sentiment phrase aligned system between a resource-rich language English and a resource-constrained language Bengali.

The scarcity in state of the art sentiment aligned translation system motivates us to perform this task. To the best of our knowledge, no previous work has been done for the English-Bengali language pair translation by considering the sentiment-aligned approach.

In our approach, sentiment expressions, sentiment holder and the corresponding objects of the holders are used to improve the phrase alignment

of the SMT system during training stage. Sentiment information is also used in the automatic post-editing of the SMT output after the decoding phase. SMT is based on a mathematical model, is most reliable and cost effective in many applications. This is one of the main reasons to choose SMT for our English-Bengali translation task. For automatic post-editing, we marked the phrases that contain sentiment expression, holders and their corresponding object. After translating the marked-up sentences, we then restructure the restructure the output according to the sentiment relations between the sentiment holder and the sentiment expression. Our approach involves the following steps:

- We first identify phrases, which contain sentiment holder, sentiment expressions and their corresponding objects.
- We aligned these phrases using word alignment provided by GIZA++.
- The aligned phrases are incorporated with the PB-SMT phrase table.
- Finally, the automatic post-editing has been carried out using the positional information of sentiment components.

The rest of the paper is organized in the following manner. Section 2 briefly elaborates the related work. Section 3 provides an overview of the dataset used in our experiments. The proposed system is described in Section 4 while Section 5 provides the system setup for the various experiments. Section 6 includes the experiments and results obtained. Finally, Section 7 concludes and provides avenues for further work.

2 Related Work

SMT systems have undergone considerable improvements over the years. Moreover, PB-SMT models (Koehn et al., 2003) outperform word-based models. The alignment template approach for PB-SMT (Och et al., 2004) allows many-to-many relations between words. A model that uses hierarchical phrases based on synchronous grammars is presented in (Chiang et al., 2005). To date there is little research on English-Bengali SMT: PB-SMT systems can be improved (Pal et al., 2011; 2013) by single tokenizing Multiword Expressions (MWEs) on both sides of the parallel corpus. Researches on alignment were mostly developed for MT tasks (Brown, 1991; Gale and

Church, 1993). A Maximum Entropy model based approach for English-Chinese NE alignment has been proposed in Feng et al. (2004), which significantly outperforms IBM Model 4 and HMM. Fung (1994) presented K-vec, an alternative alignment strategy that starts by estimating the lexicon.

Sentiment detection is the task of determining positive or negative sentiment of words, phrases, sentences and documents. The computational approach to sentiment analysis in textual data requires annotated lexicons with polarity tags (Patra et al., 2013). Research has been carried out on building sentiment or emotional corpora in English (Strapparava and Valitutti, 2004; Baccianella et al., 2010; Patra et al., 2013) and Bengali (Das and Bandyopadhyay, 2010; Das and Bandyopadhyay, 2010a). Identifying the sentiment holder is another task closely related to subjectivity detection (Kim and Hovy, 2004). Several methods have been implemented to identify the sentiment holders such as rule based methods (using dependency information) (Kolya et al., 2012) and supervised machine learning methods (Kim and Hovy, 2004; Kolya et al., 2012).

To the best of our knowledge, no prior work on improving SMT systems using aligned sentiment expressions, holders and their corresponding objects have been developed yet. There is research on creating sentiment lexica and cross-lingual sentiment identification. Automatic translation is a viable alternative for the construction of resources and tools for subjectivity or sentiment analysis in a new resource-constrained language using a resource-rich language as a starting point (Banea et al., 2008). Banea et al., (2008) generated resources for subjectivity annotation in Spanish and Romanian using English corpora. In context of Indian languages, Das et al., 2010 have developed a sentiment lexicon for Bengali Languages using an English to Bengali MT system. Similarly, a Hindi sentiment corpus has been developed using English to Hindi MT system (Balamurali et al., 2010). Hiroshi et al., (2004) developed a high-precision sentiment analysis system with low development cost, by making use of an existing transfer-based MT engine.

3 Dataset

In our experiment, an English-Bengali parallel corpus containing 23,492 parallel sentences com-

prising of 488,026 word tokens from the travel and tourism domain has been used. We randomly selected 500 sentences each for the development set and the test set from the initial parallel corpus. The rest of the sentences were used as the training corpus. The training corpus was filtered with the maximum allowable sentence length of 100 words and sentence length ratio of 1:2 (either way). The corpus has been collected from the “Development of English to Indian Languages Machine Translation (EILMT) System¹” project.

4 System Description

Initially, we identify the sentiment expressions, holders and objects from English-Bengali parallel sentences. Sentiment phrase alignment model has been developed using our existing baseline table provided by GIZA++. These aligned sentiment phrases are integrated with the state-of-the-art PB-SMT system. Finally, an automatic post editing system has been developed to correct the translation output using the textual clues identified from the sentiment components.

4.1 Sentiment expression, holder and object identification from Parallel corpus

Sentiment: Initially, sentiment expressions were not tagged with sentiment polarity. Therefore, we developed a bootstrapping method to tag the words with sentiment polarity. We have tagged the English sentiment words using the SentiWordNet 3.0 (Baccianella et al., 2010). The raw English sentences were parsed and the stems of the words were extracted using the Stanford parser². SentiWordNet examines stemmed words along with their part of speech and provides a sentiment score for each stemmed word. The sentiment of the word is judged positive, negative or neutral according to its sentiment scores. We have manually created a stop word list of around 300 words that helps us to remove the stop words from the sentences. But the words ‘not’, ‘neither’ etc. are not removed as they are valence shifters and can change the sentiment of the whole sentence. We identified 76924 and

36125 number of positive and negative words respectively.

Holder (Subject Based): Sentiment analysis involves identifying its associated holder and the event or topic. A sentiment holder is the person or organization that expresses the positive or negative sentiment towards a specific event or topic. English input sentences are parsed by the Stanford Parser to extract the dependency relations. The output is checked to identify the predicates (i.e., “*nsubj*” and “*xsubj*”), so that the *subject* related information in the “*nsubj*” and “*xsubj*” predicates are considered as probable candidates of sentiment holders.

We correlate our sentiment words with the holder using the dependency tree. For example, the sentence “*I hate chocolate but he loves it.*” has two sentiment expressions, “hate” and “love”. Here the root word and the sentiment expression is the same, i.e. “hate”. We identify that the sentiment expression, “hate” and subject “*I*” are related with “*nsubj*” relation. We conclude that “*I*” is the sentiment holder of the word “hate”. Similarly, we identify that “*he*” is the sentiment holder of word “loves”.

Example 2: **nsubj**(hate-2, I-1), **root**(ROOT-0, hate-2), **doj**(hate-2, **chocolate**-3), **nsubj**(loves-6, **he**-5), **conj_but**(hate-2, loves-6), **doj**(loves-6, **it**-7).

We have identified only 22992 number of sentiment holders, in comparison to a total of 113049 sentiment expressions.

Object: The parsed data were analyzed to identify the object of a sentence. It is found that the relations, “*doj*” and “*obj*” are considered as the probable candidates for the object. The above example sentence along with parsed output and dependency relations (example 2), the “*doj*” dependency relation includes the object. Here, “*chocolate*” and “*it*” are identified and tagged as the “*object*”.

4.2 Sentiment Phrase Alignment

In case of low-resource languages, chunking the parallel sentences (both source and target) adds more complexity in building any system. POS taggers or Chunkers might not be available for some low-resource languages. In such cases, the methodology we present below can help chunk sentences. In this paper, we propose a simple but effective

1 The EILMT project is funded by the Department of Electronics and Information Technology (DEITY), Ministry of Communications and Information Technology (MCIT), Government of India.

² <http://nlp.stanford.edu/software/lex-parser.shtml>

chunking technique. The sentence fragments are very similar with grammatical phrases or chunks. We collected the stop word lists for English as well as Bengali to implement this method (Groves and Way, 2005). We chop a sentence into several fragments whenever a stop word is encountered.

Example 3: English sentence fragmentation

“In 1905, <holder>Calcutta</holder> <expression_1>protested</expression_1> the <object_1>partition</object_1> of Bengal and <expression_2>boycotted</expression_2> all the <object_2>British Goods</object_2>.”

1. (In 1905) 2. (, Calcutta protested) 3. (the partition) 4. (of Bengal) 5. (and boycotted) 6. (all) 7. (the British Goods)

Sentiment relation

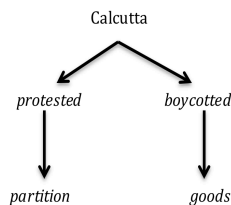


Figure 1

Phrasal sentiment relation

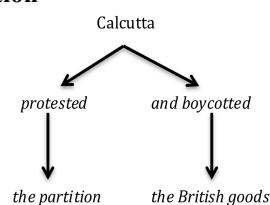


Figure 2

Bengali sentence fragmentation:

“1905 sale, Kolkata bongo vongo-r protibad janiyechilo ebong ssmosto british samogri boykot korechilo.”

Pre-processing: 1905 sale , Kolkata bongo vongo -r protibad janiyechilo ebong ssmosto british samogri boykot korechilo.

1. (1905 sale) 2. (, Kolkata bongo vongo) 3. (-r protibad janiyechilo) 4. (ebong ssmosto british samogri boykot korechilo.)

Initially, we built an English-Bengali word alignment model, which was trained with the same EILMT tourism domain parallel corpus of 22,492 sentences. Using this word alignment knowledge we aligned bilingual sentiment phrases. For establishing the alignment, we use the same phrase alignment algorithm which is used in existing state-of-the-art PB-SMT system Moses. The rest of the processes, such as scoring and phrase table creation also follow the state-of-the-art system.

4.3 Automatic Post Editing using Sentiment Knowledge

Begin The decoding process is carried out with the Moses decoder and the PB-SMT model is comput-

ed with Moses. Recall our previous example, and that after translation, the sentiment relation may interchange, so that the semantic meaning of the sentence may be the opposite of what was stated in the source. For example:

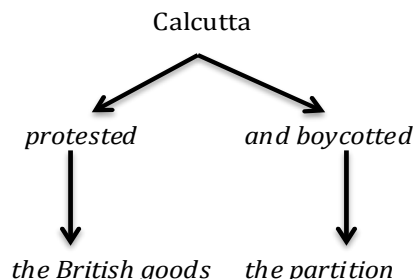


Figure 3

To correct this error in translation, we repositioned the words using our source sentiment relation knowledge obtained on the English side of the data. First, we marked the input source sentence with sentiment holder, sentiment word and their corresponding objects and measured the distance between them. The distance is then also measured on the translated sentence. If they maintained the similar distance, i.e. it ensures that they did not exchange their positions. Otherwise, we have to exchange the position of sentiment expression with their corresponding holder and object according to the distance measure. In this way, we automatically post-edited the entire translated sentences, if required.

5 System Setup

The effectiveness of the present work is demonstrated by using the standard log-linear PB-SMT model as our baseline system. For building baseline system, we use the maximum phrase length of 7 and a 5-gram language model. The other experimental settings were: GIZA++ implementation of IBM word alignment model 4 with grow-diagonal-final-and heuristics for performing word alignment and phrase-extraction (Koehn et al., 2003). The reordering model was trained msd-bidirectional (i.e. using both forward and backward models) and conditioned on both source and target languages. The reordering model is built by calculating the probabilities of the phrase pair associated with the given orientation such as monotone (m), swap(s) and discontinuous (d). We use Minimum Error Rate Training (MERT) (Och, 2003) on a held-out

development set of 500 sentences, and a target language model with Kneser-Ney smoothing (Kneser and Ney, 1995) trained with SRILM (Stolcke, 2002).

6 Experiments and Results

Our experiments have been carried out in two directions. First we improved the baseline model using the aligned sentiment phrases. Then, we automatically post-edited the translation output by using the sentiment knowledge of the source input test sentence.

The evaluation results are reported in Table 1. The evaluation was carried out using well-known automatic MT evaluation metrics: BLEU (Papineni et al., 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), and TER (Snover et al., 2006). In experiment 2, the extracted parallel sentiment phrase alignments are incorporated with the existing baseline phrase table and the resulting model performs better than the baseline system. Experiment 3 shows how post-editing the output of experiment 2 brings about further improvements.

7 Conclusions and Future Research

In this paper, we successfully illustrated how sentiment analysis can improve the translation of an English-Bengali PB-SMT system. We have also shown how sentiment knowledge is useful for automatic post-editing the MT output. In either case, we were able to improve the performance over the baseline system. Using sentiment phrase alignment we obtained a 25.73% relative improvement in BLEU over the baseline system. The automatic post-editing method results in a 33.88% relative improvement in BLEU over the baseline system. On manual inspection of the output translation we found that after incorporating sentiment phrase alignment with the baseline PB-SMT system, the output delivers better lexical selection. The post-editing method also ensures better word ordering to some extent. In the near future, we will extend the post-editing process and improve our sentiment

alignment strategies by using machine learning algorithms.

Acknowledgments

The research leading to these results has received funding from the EU FP7 Project EXPERT the People Programme (Marie Curie Actions) (Grant No. 317471) and the “Development of English to Indian Languages Machine Translation (EILMT) System - Phase II” project funded by Department of Information Technology, Government of India.

References

- Aditya Joshi, A. R. Balamurali and Pushpak Bhattacharyya. 2010. A fall-back strategy for sentiment analysis in Hindi: a case study. In *Proceedings of the 8th International Conference on Natural Language Processing*.
- Amitava Das and Sivaji Bandyopadhyay. 2010. Senti-WordNet for Indian Languages. In *Proceedings of the 8th Workshop on Asian Language Resources (ALR)*, August, pp. 56-63.
- Andreas Stolcke. 2002. SRILM-An Extensible Language Modeling Toolkit. In *Proceedings of the Intl. Conf. on Spoken Language Processing*, pp. 901–904.
- Anup K. Kolya, Dipankar Das, Asif Ekbal and Sivaji Bandyopadhyay. 2012. Roles of event actors and sentiment holders in identifying event-sentiment association. In *Proceedings of CICLing-2012*, pp. 513-525.
- Arthur P. Dempster, Nan M. Laird and Donald B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39 (1): 1–38.
- Braja G. Patra, Hiroya Takamura, Dipankar Das, Manabu Okumura and Sivaji Bandyopadhyay. 2013. Construction of Emotional Lexicon Using Potts Model. In *Proceedings of the 6th IJCNLP-2013*, Nagoya, Japan, pp. 674-679.
- Carlo Strapparava and Alessandro Valitutti. 2004. Wordnet-affect: an affective extension of wordnet. In *Proceedings of the 4th LRE*, Lisbon, pp. 1083-1086.
- Carmen Banea, Rada Mihalcea, Janyce Wiebe and Samer Hassan. 2008. Multilingual subjectivity analysis using machine translation. In *Proceedings of the EMNLP*, pp. 127-135.

Exp. No.	Experiments	BLEU	NIST	METEOR	TER
1	Baseline (B)	10.92	4.16	0.3073	75.34
2	B + Sentiment Phrase Alignment (SPA)	13.73	4.44	0.3307	72.93
3	B + SPA + Post Editing	14.62	4.44	0.3416	71.83

Table 1: Evaluation Result

- David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of 43rd Annual Meeting on Association for Computational Linguistics*, pp. 263-270.
- Declan Groves and Andy Way. 2005. Hybrid data-driven models of machine translation. *Machine Translation* 19(3-4): 301-323.
- Dipankar Das and Sivaji Bandyopadhyay. 2010. Developing Bengali WordNet Affect for Analyzing Emotion. In *International Conference on the Computer Processing of Oriental Languages*, pp. 35-40.
- Donghui Feng, Yajuan Lü and Ming Zhou. 2004. A New Approach for English-Chinese Named Entity Alignment. In *Proc. of the EMNLP*, pp. 372-379.
- Franz Josef Och and Hermann Ney. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational linguistics* 30(4): 417-449.
- George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proc. of the 2nd Human Language Technology Research*, pp. 138-145.
- Hua Wu, Haifeng Wang and Chengqing Zong. 2008. Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In *Proc. of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Manchester, UK, pp. 993-1000.
- Jaehyung Yang. 2004. Phrase chunking for efficient parsing in machine translation system. In *MICAI 2004: Advances in Artificial Intelligence*, pp. 478-487.
- Kanayama Hiroshi, Nasukawa Tetsuya and Watanabe Hideo. 2004. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th international conference on Computational Linguistics*.
- Kneser, Reinhard, and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 181-184.
- Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, pp. 311-318.
- Marine Carpuat and Mona Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. In *Proc. of HLT-NAACL*, Los Angeles, CA, pp. 242-245.
- Pascale Fung and Kenneth Ward Church. 1994. K-vec: A new approach for aligning parallel texts. In *Proceedings of the 15th Association for Computational Linguistics*.
- Peter F. Brown, Vincent J. D. Pietra, Stephen A. D. Pietra and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263-311.
- Philipp Koehn, Franz Josef Och and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of HLT-NAACL*, Edmonton, Canada, pp. 48-54.
- Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the EMNLP*, pp. 388-395.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proc. of the 45th Annual meeting of the Association for Computational Linguistics (ACL 2007)*, pp. 177-180.
- Santanu Pal, Sudip Kumar Naskar and Sivaji Bandyopadhyay. 2013. MWE Alignment in Phrase Based Statistical Machine Translation. In *Proceedings of the Machine Translation Summit XIV*, Nice, France. pp. 61-68
- Santanu Pal, Tanmoy Chkraborty and Sivaji Bandyopadhyay. 2011. Handling Multiword Expressions in Phrase-Based Statistical Machine Translation. In *Proceedings of the Machine Translation Summit XIII*, Xiamen, China. pp. 215-224.
- Satanjeev Banerjee and Alon Lavie. 2005. An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 65-72. Ann Arbor, Michigan.
- Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, pp. 1-8.
- Soo-Min Kim and Eduard Hovy. 2006. Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*. pp. 1-8. ACL.
- Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.
- Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61-74.
- William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational linguistics* 19(1): 75-102.