MWE Alignment in Phrase Based Statistical Machine Translation

Santanu Pal[*], Sudip Kumar Naskar[†] and Sivaji Bandyopadhyay[*]
[*]Department of Computer Science & Engineering
Jadavpur University, Kolkata, India
santanu.pal.ju@gmail.com, sivaji_cse_ju@yahoo.com
[†] Department of Computer & System Sciences
Visva-Bharati University, Santiniketan, India
sudip.naskar@gmail.com

Abstract

Multiword Expression (MWE) contributes to major lexical ambiguity problems for any language and poses a big challenge in statistical machine translation. This paper presents the role of MWEs in improving the performance of phrase based Statistical machine Translation (PB-SMT) system. We preprocess the parallel corpus by single tokenizing the MWEs on both sides which leads to significant improvement over baseline PB-SMT system. Automatically aligned MWEs have been incorporated into PB-SMT in two ways: indirectly, i.e., added as additional parallel training examples, and directly integrated into the word alignment model. Both the indirect and direct approaches bring some improvements in terms of system performance and the improvements are at par. For MWE alignment, we used baseline PB-SMT systems trained on the same parallel corpus in both directions. String level edit distance is used for alignment validation. We bootstrap the whole procedure to get more MWE alignments. Integration of MWE alignment into PB-SMT achieves significant improvements (7.0 BLEU points absolute, 64.1% relative improvement) over the baseline, while bootstrapping with single iteration provides further improvement (9.24 BLEU points absolute, 84.7% relative improvement) in an English—Bengali translation task.

1 Introduction

A very good quality word and phrase alignment which acquires the translation knowledge from a parallel corpus improves the performance of a Statistical machine translation (SMT) system. In this paper we handle the problem of multiword Expression (MWE) as a lexical ambiguity in phrase based statistical machine translation system. The proposed solution improves the word alignment quality.

The main problem in Machine Translation (MT) is ambiguity. Ambiguous words possess more than one meaning depending on the context they are used in. When a word is used in conjunction with other word(s), even if each of these words possesses only one meaning, they can also become ambiguous. Two common examples of this kind in English are phrasal verbs and idioms. The component words of MWE have their own separate meanings when they occur independently and the meaning of the MWE cannot always be derived from them. Examples include conjunctions ('as well as'), idioms ("keep one's fingers crossed" meaning "to hope a positive response"), phrasal verbs ("find out"), compound noun ("bus stop"), phrasal preposition ("according to") etc. MWE can be roughly defined as idiosyncratic interpretations that cross word boundaries (Sag et al., 2002). In most of the South Asian languages such as Bengali, Hindi etc., patterns such as adjective/adverb/noun +verb (conjunct verbs) or verb + verb (compound verbs) are considered as complex predicates. Morphological knowledge is compulsory to identify such complex Predicates (CPs) in Bengali, as Bengali is a morphologically rich language. Complex predicates in Bengali are: compound verbs (e.g., মেরে ফেলা [*mere phela*] 'to kill'), conjunct verbs (e.g.,

Sima'an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 61–68.

ভরসা করা [*bharsha kara*] 'to depend'), etc. Compound verbs consist of two verbs; the first verb is called the full verb, represented either as conjunctive participial form -এ [*–e*] or the infinitive form -তে [*–te*] at the surface level. The other verb called a light verb bears the inflection based on tense, aspect and person information of the subject. These light verbs are semantically lightened, polysemous and are limited into some definite candidate seeds (Paul, 2010). On the other hand, each Bengali conjunct verb consists of adjective, adverb or noun followed by a light verb.

Complex predicates are also reflected as Multi Word Expressions (MWEs) since the conventional meaning of the Light Verbs in Complex Predicates is absent (Baldwin and Kim, 2010, Sinha, 2009). The other types of predicates termed as Serial Verb (SV) (Mukherjee et al., 2006); follows the same lexical pattern like compound verb but the Full Verb and Light Verb behave as independent syntactic entities (e.g নিয়ে গেল niye gelo 'take-go').

In this experiment, we propose the improvement of word alignment quality. Our objective is to perceive the effectiveness of MWEs in word alignment by enhancing the quality of translation in the SMT system. In the present work, several types of MWEs like phrasal prepositions and Verb-object combinations, noun-noun compounds are automatically extracted on the source side while noun-noun compounds, reduplicated phrases and complex predicates are identified on the target side of the parallel corpus. We use simple rule-based and statistical approaches to identify these MWEs. We have also extracted MWEs from comparable corpora which enhance not only MWE identification quality but also provide *out-of-vocabulary* words to SMT system. This also helps to accrue some knowledge of out of domain data. Source and target language MWEs are aligned using a Hybrid technique. A well-known practice in domain adaptation in SMT (Eck et al., 2004; Wu et al., 2008) is to incorporate bilingual dictionaries to the training corpus; which affects on the instances of atomic translation pairs. The work has been carried out into three direction (i) The parallel corpus has been modified by single tokenization of MWEs, (ii) The alignment of MWEs are added in the parallel corpus as additional data to improve the word alignment as well as the phrase alignment quality and (iii) The alignment of MWE has been directly incorporated into the word alignment model. The preprocessing of the parallel corpus results in improved MT quality in terms of automatic MT evaluation metrics.

The remainder of the paper is organized as follows. Next section briefly elaborates the related work. The English-Bengali PB-SMT system is described in Section 3. Section 4 states the tools and resources used for the various experiments. Section 5 includes the results obtained, together with some analysis. Section 6 concludes and provides avenues for further work.

2 Related Work

Venkatapathy and Joshi (2006) reported a discriminative approach to use the compositionality information of verb-based multi-word expressions in order to improve the word alignment quality. A log likelihood ratio based hierarchical reducing algorithm to automatically extract bilingual MWEs has been described in Ren et al. (2009). They examined the usefulness of these bilingual MWEs in SMT by integrating bilingual MWEs into the Moses decoder (Koehn et al., 2007). They observed the highest improvement with an additional feature that identifies whether or not a bilingual phrase contains bilingual MWEs. This approach was generalized in Carpuat and Diab (2010) who replaced the binary feature by a count feature representing the number of MWEs in the source language phrase. A hybrid approach to identify MWEs from the English-French parallel corpus proposed by (Bouamor et al., 2012a), they aligned only many to many correspondences and deals with highly correlated MWE in a sentence pair, those are then integrated into the MOSES SMT System (Bouamor et al., 2012b). MWEs in SMT for Verbmobil corpus has also been proposed by (Lambert et al., 2005), the performance of the system evaluated in terms of alignment and translation quality.

Instinctively, MWEs on the source and the target sides should be both aligned in the parallel corpus and translated as a whole. However, the constituents of an MWE are identified and aligned as parts of consecutive phrases in the state-of-the-art PB-SMT systems, since PB-SMT (or any other approaches to SMT) does not generally treat MWEs as special tokens. Another problem with SMT systems is the wrong translation of verb phrases. Sometimes verb phrases are deleted in the output sentence. Moreover, the

words inside verb phrases are generally not aligned one-to-one; the alignments of the words inside source and target verb phrases are mostly many-to-many, particularly so for the English—Bengali language pair. These are the motivations behind considering MWEs like NEs, reduplicated phrases, prepositional phrase and compound verbs for special treatment in this work.

By converting the MWEs into single tokens, we make sure that PB-SMT also treats them as a whole. The first objective of the present work is to see how single tokenization of MWEs on both the sides affects the overall MT quality. The second objective is to see whether prior automatic alignment of single-tokenized MWEs can improve the machine translation quality. The third objective is to see whether automatic alignment of single-tokenized MWEs incorporated directly into the word alignment model can bring any further improvement in the overall performance of the MT system.

We carried out the experiments on English—Bengali translation task. Bengali shows high morphological richness at lexical level. Language resources in Bengali are not widely available. Furthermore, this is the first time when the identification of MWEs in Bengali language is used to enhance the performance of an English-Bengali Machine Translation System.

3 System Description

3.1 Preprocessing of the parallel corpus

We considered several types of multi-word expressions: noun-noun MWEs, reduplicated phrases, complex predicates, phrasal prepositions, and verb-object combination. For the identification of complex predicates, we adopted a similar technique as reported in (Das et al., 2010). There are no frequent occurrences of reduplicated phrases in the English Corpus (Chakraborty and Bandyopadhyay, 2010) in comparison with the Bengali corpus, so this plays very crucial role in machine translation as they occur with high frequently in the Bengali corpus.

Once the MWEs are identified, they are converted into single-tokens by replacing the spaces with underscores (‘_’) so that we can establish 1-to-1 alignments between the source and target MWEs.

3.2 MWE Identification

Noun-Noun MWE Identification: When two or more nouns are united together to form a solo phrase such as ‘bed room’ or ‘dining table’ (Baldwin and Kim, 2010), these are termed as compound nouns or nominal compounds. Compound noun MWEs can be defined as a lexical unit made up of two or sometimes more elements, in different contexts, each of which can function as a lexeme independent of the others(s). There are some phonological and/or grammatical isolation from normal syntactic usage shown in compound noun MWEs. A number of techniques have already been applied for MWE identification. In this experiment, we have followed Point-wise Mutual Information (PMI) (Church et al.1990), Log-likelihood Ratio (LLR) (Dunning 1993) and Phi-coefficient, Co-occurrence measurement and significance function (Agarwal et al. 2004) measures. Finally, a system combination model has been developed which gives a normalized weighted combination score to each of the extracted MWEs. A predefined cut-off score (above 70%) has been considered and the candidates having scores above the threshold value have been considered as MWEs. Various types of MWEs are recognized by our system such as phrasal verbs (e.g., stubbed out), noun phrases (e.g., running train), proper names (e.g., Mahatma Gandhi) etc. Similar method has been followed to identify the other MWEs.

Identification of Reduplication: In all languages, the repetition of noun, pronoun, adjective and verb may be at the expression level or contents or semantic level. In this experiment we have considered only expression level reduplication on the Bengali corpus. The expression-level Bengali reduplications are further classified into five fine-grained subcategories (Chakraborty and Bandyopadhyay, 2010): (i) Onomatopoeic expressions: The sound sequence of the word denotes the particular meaning of the form (খট খট, *khat khat*, knock knock), (ii) Complete Reduplication: The individual words carry certain meaning, and they are repeated (বড় বড়, *bara-bara*, big big), (iii) Partial Reduplication: Only one of the words is meaningful, while other is constructed by partially reduplicating the first word (ঠাকুর ঠুকুর, *thakur-thukur*, God), (iv) Semantic Reduplication: The paired members are semantically related such as synonymy (মাথা মুন্ডু,

matha-mundu, head), antonymous (দিন-রাত, *din-rat*, day and night) etc. and (v) Correlative Reduplication: The corresponding correlative words is used just preceding the main verb (মারামারি, *maramari*, fighting). The present work tries to cover almost all the above mentioned types. We have used a simple rule-based approach (Chakraborty and Bandyopadhyay, 2010) in the present work to identify reduplication in corpus.

3.3 MWE Extraction from Comparable Corpora

As the parallel training set for our experiments was relatively small, we collected comparable corpora from Wikipedia. Wikipedia1 is an online collaborative encyclopedia available in a wide variety of languages. English Wikipedia is the largest in volume with millions of articles; there are many language editions with at least 100,000 articles. Wikipedia use "interwiki" linking facility to link articles on the same topic in different languages. Wikipedia is an enormously useful resource for extracting parallel resources as the documents in different languages are already aligned. We listed the named entities (NEs) from the training data and aligned them through transliteration following the approach of (Pal et al., 2010). From this parallel NE list we search information about individual NEs in Wikipedia for both the source and the target languages. Wikipedia provides document-level aligned comparable corpora. We identify MWEs from both sides of the comparable corpora following the method described in section 3.2 and align them following the procedure described in section 3.4.

3.4 Automatic Alignment of MWEs

The initial English−Bengali parallel corpus is cleaned and filtered using an automatic process. On the existing parallel corpus very few types of MWEs have been identified. As the MWE identification method follows a statistical method, we add comparable corpus with our training data to get more collocation value. Using comparable corpora we have extracted much more MWEs from the training data. An English−Bengali PB-SMT and a Bengali−English PB-SMT system have been developed to translate English MWEs and Bengali MWEs respectively. The English MWEs are translated and validated against the target Bengali MWEs extracted from the Bengali corpus Bengali corpus and are saved as a separate list. Similarly extracted Bengali MWEs are also translated and validated using source English MWEs extracted from the English corpus. The extracted MWEs from the comparable corpora are also aligned following similar methods as described earlier and produce an MWE dictionary. The MWE alignment system architecture is depicted in Figure 1.

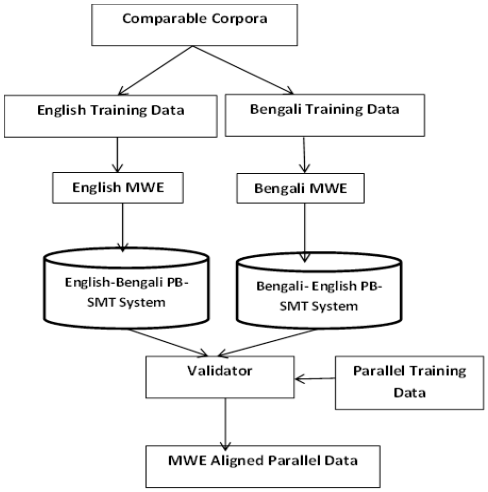

Figure 1. MWE Alignment System Architecture

3.5 MWE Alignment Validation

The validation process uses a fuzzy matching technique for validating the MWE alignments between translated MWEs and training text. A closely matching string is identified from the corresponding parallel text of the extracted MWEs. To find the closest match, we used a fuzzy matching score based on character level edit distance metric (Wagner and Fischer, 1974). The fuzzy matching score between two strings is defined as in equation 1. The closest matching string from the parallel sentence is associated with the corresponding MWE.

For two string m_i and m_j the fuzzy matching score is:

$$Score(m_i, m_j) = \frac{1-ED(m_i,m_j)}{Max\,(|m_i|,|m_j|)} \quad (1)$$

where $|m_i|$ denotes the length (in characters) of the MWE m_i and ED denotes edit distance between two string m_i and m_j .

After retrieving the closest fuzzy matching strings for all MWEs, we prepare a MWE-level parallel corpus. These parallel MWEs are added

[1] http://www.wikipedia.org/.

with the parallel training corpus as additional training data.

3.6 Incorporating Alignment directly into the word alignment Model

Aligned bilingual MWEs have been incorporated directly into the word alignment model by updating the word alignment table. The word alignment table is updated by looking up this bilingual MWE dictionary which was extracted from the training corpus. The probability is normalized in both the source−target and target−source lexical file accordingly. The lexical file is generated during the training phrase. The phrase extraction step is continued after updating the lexical files. The phrase table creation has been carried out following the state-of-art method.

4 Tools and Resources

A sentence-aligned English-Bengali parallel corpus containing 23,492 parallel sentences from the travel and tourism domain has been used in the present work. The corpus has been collected from the consortium-mode project "Development of English to Indian Languages Machine Translation (EILMT) System[2]". The Stanford Parser[3], Stanford NER, CRF chunker[4] (Xuan-Hieu Phan, 2006) and the Wordnet 3.0[5] have been used for identifying complex predicates in the source English side of the parallel corpus.

The sentences on the target side (Bengali) are POS-tagged by using the tools obtained from the consortium mode project "Development of Indian Language to Indian Language Machine Translation (IL-ILMT) System[6]". NEs in Bengali are identified using the NER system of Ekbal and Bandyopadhyay (2008). We have used the Stanford Parser and the Bengali NER.

The effectiveness of the MWE-aligned parallel corpus is demonstrated by using the standard log-linear PB-SMT model as our baseline system: GIZA++ implementation of IBM word alignment model 4, phrase-extraction heuristics described in (Koehn et al., 2003), minimum-error-rate training (Och, 2003) on a held-out development set, target language model trained using SRILM toolkit (Stolcke, 2002) with Kneser-Ney smoothing (Kneser and Ney, 1995) and the Moses decoder (Koehn et al., 2007) have been used in the present study.

5 Experiments and Evaluations

We have randomly identified 500 sentences each for the development set and the test set from the initial parallel corpus. The rest are considered 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). Finally the training corpus contained 22,492 sentences. In addition to the target side of the parallel corpus, a monolingual Bengali corpus containing 488,026 words from the tourism domain was used for the target language model. We experimented with different n-gram settings for the language model and the maximum phrase length and found that a 4-gram language model and a maximum phrase length of 7 produce the optimum baseline result on both the development and the test set. We carried out the rest of the experiments using these settings.

The system continues with the various preprocessing of the corpus. Our hypothesis focuses mainly on the theme that as much MWEs are identified and aligned properly as possible to show the improvement of the system performance in terms of translation quality. Table 1 shows the MWE statistics of the parallel training corpus. It can be observed from Table 1 that NEs and complex predicates occur with high frequency in both sides compared to other types of MWEs. It suggests that prior alignment of the NEs plays a role in improving the system performance. Table 1 also reports that the use of comparable corpora with the training corpus improves the performance of MWE extraction to some extent.

Of all the MWEs in the training and development sets, the translation-based alignment process was able to establish alignments of 4,971 CPs, 15 reduplicated words and 676 Noun-noun compounds, but during second iteration process, the new translation model was able to aligned 7,019 CPs and 1,223 of noun-noun compounds.

The experiments have been carried out in various experimental settings: (i) single tokenization

[2] The EILMT project is funded by the Department of Electronics and Information Technology (DEITY), Ministry of Communications and Information Technology (MCIT), Government of India.
[3] http://nlp.stanford.edu/software/lex-parser.shtml
[4] http://crfchunker.sourceforge.net/
[5] http://wordnet.princeton.edu/
[6] The IL-ILMT project is funded by the Department of Electronics and Information Technology (DEITY), Ministry of Communications and Information Technology (MCIT), Government of India.

of MWEs on both sides in the parallel corpus, (ii) single tokenized MWEs added with the parallel training data, (iii) single tokenized MWEs directly integrated into the word alignment model, and finally, (iv) bootstrapping with single iteration using the experimental setup (ii) and (iii) to examine how the parallel MWE alignment set can be increased. Extrinsic evaluation was carried out on the MT quality using the well-known automatic MT evaluation metrics: BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) and the evaluation results are reported in Table 2. By considering single tokenization (experiment 2), the system achieves performance improvement to some extent. Use of comparable corpora (experiment 3) improves the MWE identification performance which in turn improves the translation quality.

Training set	**English**		**Bengali**	
	T	U	T	U
CPs	8142	3889	20174	7154
reduplicated word	55	15	185	150
Noun-noun compound	892	711	489	300
Noun-noun compound with Comparable corpora	1792	981	889	700
Phrasal preposition	1782	1379	-	-
Verb-object combination	231	145	-	-
Phrasal verb	549	532	-	-
Total NE words	29931	12273	18107	12106

Table 1.MWE Statistics. (T - Total occurrence, U – Unique, CP – Complex Predicates, NE- Named Entities)

The rest of the experiments have been carried out by upgrading the experiment 3's model. The performance improves substantially when we use the aligned MWEs as parallel examples (experiment 4) or incorporate the MWE alignment information directly into the word alignment model (experiment 5). It is to be noticed that incorporation of parallel MWE information indirectly (i.e., experiment 4) and directly (i.e., experiment 5) into PB-SMT both result is almost similar improvement. Experiment 6 and 7 represent the bootstrapping approach to MWE alignment, which follows the similar experimental setup as described in (ii) and (iii). The bootstrapping approach to MWE alignment also provides significant improvement. If we continue further iteration, we can expect further improvements; however with bootstrapping approach, the improvements tend to diminish gradually.

Our best system without bootstrapping provided 7.0 BLEU points (64.1% relative) improvement over the baseline system. While using single iteration the performance increased significantly (9.25 BLEU points, 84.7% relative) over the baseline system.

We compared the translation outputs produced by our best system against the baseline outputs for a small subset of the test data. We found that our system results in more accurate lexical choices particularly for MWEs.

Experiments	**No.**	**BLEU**	**NIST**
Baseline	1	10.92	4.13
Baseline With Single tokenize MWE (extracted from training set)	2	13.03	4.34
Baseline With Single tokenize MWEs (extracted from training set with the help of comparable corpora)	3	13.81	4.44
Exp-3+MWE Alignment	4	17.82	4.49
Exp-3 + MWE Alignment incorporated directly into Word Alignment model	5	17.92	4.49
Exp-4 + Bootstrap MWE alignment with single iteration'†'	6	19.97	4.67
Exp-5 + Bootstrap MWE Alignment with single iteration'†'	7	**20.17**	**4.68**

Table 2.Evaluation results for different experimental setups. (The '†' marked systems produce statistically significant improvements on BLEU over the baseline system)

6 Conclusion and Future Work

The study shows how effective pre-processing of MWEs in the parallel corpus, their alignment and integration (directly or indirectly) into PB-SMT can improve the system performance. For scarce resource language pair, this approach can help to

improve the state-of-art machine translation quality. Our approach also shows that knowledge can be acquired from external resources like comparable corpora. Automatic prior alignment of MWEs and MWE aligned data integrated directly into the word alignment model improve the system performance significantly, while bootstrapping with single iteration provides further gains.

For future work, we will carry out experiments with more resources acquired from comparable corpora. We would also investigate into whether this approach can bring improvements of similar magnitude for larger training data.

Acknowledgement

The work has been carried out with support from the project "Development of English to Indian Languages Machine Translation (EILMT) System - Phase II" funded by Department of Information Technology, Government of India.

References

Agarwal, Aswini, Biswajit Ray, Monojit Choudhury, Sudeshna Sarkar and Anupam Basu. 2004. Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenario. *In Proc. of International Conference on Natural Language Processing (ICON)*, pp. 165-174.

Baldwin, Timothy and Su Nam Kim (2010) *Multiword Expressions*, in Nitin Indurkhya and Fred J. Damerau (eds.) *Handbook of Natural Language Processing*, Second Edition, CRC Press, Boca Raton, USA, pp. 267—292.

Banerjee, Satanjeev, and Alon Lavie. 2005. An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *In proceedings of the ACL-2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pp. 65-72. Ann Arbor, Michigan., pp. 65-72.

Bouamor, D.Semmar, N. Zweigenbeaum, P., 2012a, Automatic Construction of a MultiWord Expressions Bilingual Lexicon: A Statistical Machine Translation Evaluation Perspective, Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING 2012, Mumbai, December 2012, pp. 95–108.

Bouamor, D.Semmar, N. Zweigenbeaum, P., 2012b, Identifying bilingual Multi-Word Expressions for Statistical Machine Translation, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12),LREC 2012,Istanbul, Turkey, May 2012, pp.674-679

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. *The mathematics of statistical machine translation: parameter estimation. Computational Linguistics*, 19(2):263-311.

Carpuat, Marine, and Mona Diab. 2010. Task-based Evaluation of Multiword Expressions: a Pilot Study in Statistical Machine Translation. *In Proc. of Human Language Technology conference and the North American Chapter of the Association for Computational Linguistics conference (HLT-NAACL 2010)*, Los Angeles, CA.

Chakraborty, Tanmoy and Sivaji Bandyopadhyay. 2010. Identification of Reduplication in Bengali Corpus and their Semantic Analysis: A Rule Based Approach. *In proc. of the 23rd International Conference on Computational Linguistics (COLING 2010), Workshop on Multiword Expressions: from Theory to Applications (MWE 2010)*. Beijing, China.

Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16(1) (1990) 22-29.

Das, Dipankar, Santanu Pal, Tapabrata Mondal, Tanmoy Chakraborty, Sivaji Bandyopadhyay .2010. Automatic Extraction of Complex Predicates in Bengali, *In proc. of the workshop on Multiword expression: from theory to application (MWE-2010)*, The 23rd International conference of computational linguistics (Coling 2010),Beijing, Chaina, pp. 37-46.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)* 39 (1): 1–38.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, *In Proc. of the Second International Conference on Human Language Technology Research (HLT-2002)*, San Diego, CA, pp. 128-132.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. Computational Linguistics 19.pp. 61–74.

Eck, Matthias, Stephan Vogel, and Alex Waibel. 2004. Improving statistical machine translation in the medical domain using the Unified Medical Language System. *In Proc. of the 20th International Conference on Computational Linguistics (COLING 2004)*, Geneva, Switzerland, pp. 792-798.

Ekbal, Asif, and Sivaji Bandyopadhyay. 2008. Maximum Entropy Approach for Named Entity Recognition in Indian Languages. *International Journal*

for Computer Processing of Languages (IJCPOL), Vol. 21 (3), 205-237.

Kneser, Reinhard, and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. *In Proc. of the IEEE Internation Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, pp. 181–184. Detroit, MI.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. *In Proc. of HLT-NAACL 2003: conference combining Human Language Technology conference series and the North American Chapter of the Association for Computational Linguistics conference series*, Edmonton, Canada, pp. 48-54.

Koehn, Philipp, 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): Proc. of demo and poster sessions*, Prague, Czech Republic, pp. 177-180.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In EMNLP-2004: *Proc. of the 2004 Conference on Empirical Methods in Natural Language Processing*, 25-26 July 2004, Barcelona, Spain, pp 388-395.

Lambert, Patrik and Rafael Banchs. 2005. Data Inferred Multi-word Expressions for Statistical Machine Translation. *In Proc. of Machine Translation Summit X*, Phuket, Thailand, pp. 396-403.

Mukherjee, Amitabha, Soni Ankit and Raina Achla M. 2006. Detecting Complex Predicates in Hindi using POS Projection across Parallel Corpora. *Multiword Expressions: Identifying and Exploiting Underlying Properties Association for Computational Linguistics*, pp. 28–35, Sydney.

Och, Franz J. 2003. Minimum error rate training in statistical machine translation. *In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan, pp. 160-167.

Pal, Santanu, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay and Andy Way.2010. Handling Named Entities and Compound Verbs in Phrase-Based Statistical Machine Translation, *In proc. of the workshop on Multiword expression: from theory to application (MWE-2010)*, The 23rd International conference of computational linguistics (Coling 2010),Beijing, Chaina, pp. 46-54.

Papineni, Kishore, 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 (ACL-2002)*, Philadelphia, PA, pp. 311-318.

Paul, Soma. 2010. Representing Compound Verbs in Indo WordNet. *Golbal Wordnet Conference-2010*, pp. 84-91.

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 2009. Improving statistical machine translation using domain bilingual multiword expressions. *In Proc. of the 2009 Workshop on Multiword Expressions, ACL-IJCNLP 2009*, Suntec, Singapore, pp. 47-54.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, Mexico City, Mexico, pp. 1–15.

Sinha, R. Mahesh, K. 2009. Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus. *Multi Word Expression Workshop, Association of Computational Linguistics-International Joint Conference on Natural Language Processing-2009*, pp. 40-46, Singapore.

Stolcke, A. SRILM—An Extensible Language Modeling Toolkit. Proc. Intl. Conf. on Spoken Language Processing, vol. 2, pp. 901–904, Denver (2002).

Tanaka, Takaaki and Timothy Baldwin. 2003. Noun-Noun Compound Machine Translation: A Feasibility Study on Shallow Processing. In *Proc. of the Association for Computational Linguistics- 2003, Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 17–24.

Venkatapathy, Sriram, and Aravind K. Joshi. 2006. Using information about multi-word expressions for the word-alignment task. *In Proc. of Coling-ACL 2006: Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, Sydney*, pp. 20-27.

Wagner, R. and M. Fischer. 1974. The String-to-String Correction Problem. *Journal of the Association for Computing Machinery*, 21:168–173.

Wu, Hua 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.

Xuan-Hieu Phan, "CRFChunker: CRF English Phrase Chunker", http://crfchunker.sourceforge.net/, 2006.