Bootstrapping Method for Chunk Alignment in Phrase Based SMT

Santanu Pal

Department of Computer Science and Engineering
Jadavpur University
santanu.pal.ju@gmail.com

Sivaji Bandyopadhyay

Department of Computer Science and Engineering
Jadavpur University
sivaji_cse@yahoo.com

Abstract

The processing of parallel corpus plays very crucial role for improving the overall performance in Phrase Based Statistical Machine Translation systems (PB-SMT). In this paper the automatic alignments of different kind of chunks have been studied that boosts up the word alignment as well as the machine translation quality. Single-tokenization of Noun-noun MWEs, phrasal preposition (source side only) and reduplicated phrases (target side only) and the alignment of named entities and complex predicates provide the best SMT model for bootstrapping. Automatic bootstrapping on the alignment of various chunks makes significant gains over the previous best English-Bengali PB-SMT system. The source chunks are translated into the target language using the PB-SMT system and the translated chunks are compared with the original target chunk. The aligned chunks increase the size of the parallel corpus. The processes are run in a bootstrapping manner until all the source chunks have been aligned with the target chunks or no new chunk alignment is identified by the bootstrapping process. The proposed system achieves significant improvements (2.25 BLEU over the best System and 8.63 BLEU points absolute over the baseline system, 98.74% relative improvement over the baseline system) on an English- Bengali translation task.

1 Introduction

The objective of the present research work is to analyze effects of chunk alignment in English – Bengali parallel corpus in a Phrase Based Statistical Machine Translation system. The initial sentence level aligned English-Bengali corpus is cleaned and filtered using a semi-automatic process. More effective chunk level alignments are carried out by bootstrapping on the training corpus to the PB-SMT system.

The objective in the present task is to align the chunks in a bootstrapping manner using a Single tokenized MWE aligned SMT model and then modifying the model by inserting the aligned chunks to the parallel corpus after each iteration of the bootstrapping process, thereby enhancing the performance of the SMT system. In turn, this method deals with the many-to-many word alignments in the parallel corpus. Several types of MWEs like phrasal prepositions and Verb-object combinations are automatically identified on the source side while named-entities and complex predicates are identified on both sides of the parallel corpus. In the target side only, identification of the Noun-noun MWEs and reduplicated phrases are carried out. Simple rule-based and statistical approaches have been used to identify these MWEs. The parallel corpus is modified by considering the MWEs as single tokens. Source and target language NEs are aligned using a statistical transliteration technique. These automatically aligned NEs and Complex predicates are treated as translation examples, i.e., as additional entries in the phrase table (Pal.et al 2010, 2011). Using this augmented phrase table each individual source chunk is translated into the target chunk and then validated with the target chunks on the target side. The validated source-target chunks are con-

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 93–100,
Avignon, France, April 23 - 27 2012.

sidered as further parallel examples, which in effect are instances of atomic translation pairs to the parallel corpus. This is a well-known practice in domain adaptation in SMT (Eck et al., 2004; Wu et al., 2008). 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. Section 2 briefly elaborates the related work. The PB-SMT system is described in Section 3. The resources used in the present work are described in Section 4. The various experiments carried out and the corresponding evaluation results have been reported in Section 5. The conclusions are drawn in Section 6 along with future work roadmap.

2 Related work

A multi lingual filtering algorithm generates bilingual chunk alignment from Chinese-English parallel corpus (Zhou.et al, 2004). The algorithm has three steps, first, the most frequent bilingual chunks are extracted from the parallel corpus, second, a clustering algorithm has been used for combining chunks which are participating for alignment and finally one English chunk is generated corresponding to a Chinese chunk by analyzing the highest co-occurrences of English chunks. Bilingual knowledge can be extracted using chunk alignment (Zhou.et al, 2004). The alignment strategies include the comparison of dependency relations between source and target sentences. The dependency related candidates are then compared with the bilingual dictionary and finally the chunk is aligned using the extracted dependency related words. Ma.et al. (2007) simplified the task of automatic word alignment as several consecutive words together correspond to a single word in the opposite language by using the word aligner itself, i.e., by bootstrapping on its output. Zhu and Chang (2008) extracted a dictionary from the aligned corpus, used the dictionary to re-align the corpus and then extracted the new dictionary from the new alignment result. The process goes on until the threshold is reached.

An automatic extraction of bilingual MWEs is carried out by Ren et al. (2009), using a log likelihood ratio based hierarchical reducing algorithm to investigate the usefulness of bilingual MWEs in SMT by integrating bilingual MWEs into the Moses decoder (Koehn et al., 2007). The system has 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) where the binary feature is replaced by a count feature which is representing the number of MWEs in the source language phrase.

MWEs on the source and the target sides should be both aligned in the parallel corpus and translated as a whole. However, in the state-of-the-art PB-SMT systems, the constituents of an MWE are marked and aligned as parts of consecutive phrases, 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 some phrases. Sometimes some phrases are not found in the output sentence. Moreover, the source and target phrases are mostly many-to-many, particularly so for the English—Bengali language pair. The main objective of the present work is to see whether prior automatic alignment of chunks can bring any improvement in the overall performance of the MT system.

3 PB-SMT System Description

The system follows three steps; the first step is prepared an SMT system with improved word alignment that produces a best SMT model for bootstrapping. And the second step is produced a chunk level parallel corpus by using the best SMT model. These chunk level parallel corpuses are added with the training corpus to generate the new SMT model in first iteration. And finally the whole process repeats to achieve better chunk level alignments as well as the better SMT model.

3.1 SMT System with improved Word Alignment

The initial English-Bengali parallel corpus is cleaned and filtered using a semi-automatic process. Complex predicates are first extracted on both sides of the parallel corpus. The analysis and identification of various complex predicates like, compound verbs (*Verb + Verb*), conjunct verbs (*Noun /Adjective/Adverb + Verb*) and serial verbs (*Verb + Verb + Verb*) in Bengali are done following the strategy in Das.et al. (2010).

Named-Entities and complex predicates are aligned following a similar technique as reported in Pal.et al (2011). Reduplicated phrases do not occur very frequently in the English corpus; some of them (like correlatives, semantic reduplications) are not found in English (Chakraborty

and Bandyopadhyay, 2010). But reduplication plays a crucial role on the target Bengali side as they occur with high frequency. These reduplicated phrases are considered as a single-token so that they may map to a single word on the source side. Phrasal prepositions and verb object combinations are also treated as single tokens. Once the compound verbs and the NEs are identified on both sides of the parallel corpus, they are assembled into single tokens. When converting these MWEs into single tokens, the spaces are replaced with underscores ('_'). Since there are already some hyphenated words in the corpus, hyphenation is not used for this purpose. Besides, the use of a special word separator (underscore in this case) facilitates the job of deciding which single-token MWEs to be de-tokenized into its constituent words, before evaluation.

3.1.1 MWE Identification on Source Side

The UCREL1 Semantic analysis System (USAS) developed by Lancaster University (Rayson.et al, 2004) has been adopted for MWE identification. The USAS is a software tool for the automatic semantic analysis of English spoken and written data. Various types of Multi-Word Units (MWU) that are identified by the USAS software include: verb-object combinations (e.g. stubbed out), noun phrases (e.g. riding boots), proper names (e.g. United States of America), true idioms (e.g. living the life of Riley) etc. In English, Noun-Noun (NN) compounds, i.e., noun phrases occur with high frequency and high lexical and semantic variability (Tanaka.et al, 2003). The USAS software has a reported precision value of 91%.

3.1.2 MWE Identification on Target Side

Compound nouns are identified on the target side. Compound nouns are nominal compounds where two or more nouns are combined to form a single phrase such as *'golf club'* or *'computer science department'* (Baldwin.et al, 2010). Each element in a compound noun can function as a lexeme in independent of the other lexemes in different context. The system uses Point-wise Mutual Information (PMI), Log-likelihood Ratio (LLR) and Phi-coefficient, Co-occurrence measurement and Significance function (Agarwal.et al, 2004) measures for identification of compound nouns. Final evaluation has been carried out by combining the results of all the methods. A predefined cut-off score has been considered and the candidates having scores above the threshold value have been considered as MWEs.

The repetition of noun, pronoun, adjective and verb are generally classified as two categories: repetition at the (a) expression level and at the (b) contents or semantic level. In case of Bengali, The expression-level reduplication are classified into five fine-grained subcategories: (i) Onomatopoeic expressions (***khat khat***, knock knock), (ii) Complete Reduplication (***bara-bara***, big big), (iii) Partial Reduplication (***thakur-thukur***, God), (iv) Semantic Reduplication (***matha-mundu***, head) and (v) Correlative Reduplication (***maramari***, fighting).

For identifying reduplications, simple rules and morphological properties at lexical level have been used (Chakraborty and Bandyopadhyay, 2010). The Bengali monolingual dictionary has been used for identification of semantic reduplications.

An NE and Complex Predicates parallel corpus is created by extracting the source and the target (single token) NEs from the NE-tagged parallel corpus and aligning the NEs using the strategies as applied in (Pal.et al, 2010, 2011).

3.1.3 Verb Chunk / Complex Predicate Alignment

Initially, it is assumed that all the members of the English verb chunk in an aligned sentence pair are aligned with the members of the Bengali complex predicates. Verb chunks are aligned using a statistical aligner. A pattern generator extracts patterns from the source and the target side based on the correct alignment list. The root form of the main verb, auxiliary verb present in the verb chunk and the associated tense, aspect and modality information are extracted for the source side token. Similarly, root form of the Bengali verb and the associated vibhakti (inflection) are identified on the target side token. Similar patterns are extracted for each alignment in the doubtful alignment list.

Each pattern alignment for the entries in the doubtful alignment list is checked with the patterns identified in the correct alignment list. If both the source and the target side patterns for a doubtful alignment match with the source and the target side patterns of a correct alignment, then the doubtful alignment is considered as a correct one.

The doubtful alignment list is checked again to look for a single doubtful alignment for a sentence pair. Such doubtful alignments are considered as correct alignment.

[1] http://www.comp.lancs.ac.uk/ucrel

The above alignment list as well as NE aligned lists are added with the parallel corpus for creating the SMT model for chunk alignment. The system has reported 15.12 BLEU score for test corpus and 6.38 (73% relative) point improvement over the baseline system (Pal.et al, 2011).

3.2 Automatic chunk alignment

3.2.1 Source chunk extraction

The source corpus is preprocessed after identifying the MWEs using the UCREL tool and single tokenizing the extracted MWEs. The source sentences of the parallel corpus have been parsed using Stanford POS tagger and then the chunks of the sentences are extracted using CRF chunker[2] The CRF chunker detects the chunk boundaries of noun, verb, adjective, adverb and prepositional chunks from the sentences. After detection of the individual chunks by the CRF chunker, the boundary of the prepositional phrase chunks are expanded by examining the series of noun chunks separated by conjunctions such as '*comma*', '*and*' etc. or a single noun chunk followed by a preposition. For each individual chunk, the head words are identified. A synonymous bag of words is generated for each head word. These bags of words produce more alternative chunks which are decoded using the best SMT based system (Section 3.1). Additional translated target chunks for a single source chunk are generated.

CRF Chunker output

bodies/NNS/B-NP of/IN/B-PP all/DT/B-NP ages/NNS/I-NP ,/,/O colors/NNS/I-NP and/CC/O sizes/NNS/I-NP don/VB/B-VP the/DT/B-NP very/JJ/I-NP minimum/NN/I-NP in/IN/B-PP beachwear/NN/B-NP and/CC/O idle/VB/B-VP away/RP/B-PRT the/DT/B-NP days/NNS/I-NP on/IN/B-PP the/DT/B-NP sun/NN/I-NP kissed/VBN/I-NP copacabana/NN/I-NP and/CC/O ipanema/NN/I-NP beaches/NNS/I-NP ././O

Noun chunk Expansion and boundary detection

(bodies/NNS/B-NP) (of/IN/B-PP) (all/DT/B-NP ages/NNS/I-NP ,/,/I-NP colors/NNS/I-NP and/CC/I-NP sizes/NNS/I-NP) (don/VB/B-VP) (the/DT/B-NP very/JJ/I-NP minimum/NN/I-NP) (in/IN/B-PP) (beachwear/NN/B-NP) (and/CC/B-O) (idle/VB/B-VP) (away/RP/B-PRT) (the/DT/B-NP days/NNS/I-NP) (on/IN/B-PP) (the/DT/B-NP sun/NN/I-NP kissed/VBN/I-NP copacabana/NN/I-NP and/CC/I-NP ipanema/NN/I-NP beaches/NNS/I-NP) (././B-O)

Prepositional phrase expansion and extraction

bodies
of all ages , colors and sizes
don
the very minimum
in beachwear
and
idle
away
the days
on the sun kissed copacabana and ipanema beaches

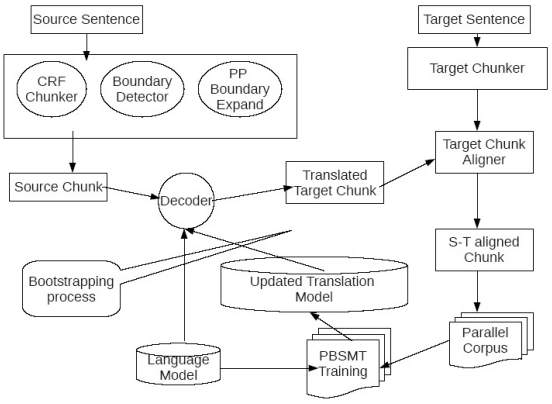

Figure 1.System architecture of the Automatic chunk alignment model

3.2.2 Target chunk extraction

The target side of the parallel corpus is cleaned and parsed using the shallow parser developed by the consortia mode project "Development of Indian Language to Indian Language Machine Translation (IL-ILMT) System Phase II" funded by Department of Information Technology, Government of India. The individual chunks are extracted from the parsed output. The individual chunk boundary is expanded if any noun chunk contains only single word and several noun chunks occur consecutively. The content of the individual chunks are examined by checking their POS categories. At the time of boundary expansion, if the system detects other POS category words except noun or conjunction then the expansion process stops immediately and new chunk boundary beginning is identified. The IL-ILMT system generates the head word for each individual chunk. The chunks for each sentence are stored in a separate list. This list is used as a

[2] http://crfchunker.sourceforge.net/

validation resource for validate the output of the statistical chunk aligner.

3.2.3 Source-Target chunk Alignment

The extracted source chunks are translated using the generated SMT model. The translated chunks as well as their alternatives are validated with the original target chunk. During validation checking, if any match is found between the translated chunk and the target chunk then the source chunk is directly aligned with the original target chunk. Otherwise, the source chunk is ignored in the current iteration for any possible alignment. The source chunk will be considered in the next alignment. After the current iteration is completed, two lists are produced: a chunk level alignment list and an unaligned source chunk list. The produced alignment lists are added with the parallel corpus as the additional training corpus to produce new SMT model for the next iteration process. The next iteration process translates the source chunks that are in the unaligned list produced by the previous iteration. This process continues until the unaligned source chunk list is empty or no further alignment is identified.

3.2.4 Source-Target chunk Validation

The translated target chunks are validated with the original target list of the same sentence. The extracted noun, verb, adjective, adverb and prepositional chunks of the source side may not have a one to one correspondence with the target side except for the verb chunk. There is no concept of prepositional chunks on the target side. Some time adjective or adverb chunks may be treated as noun chunk on the target side. So, chunk level validation for individual categories of chunks is not possible. Source side verb chunks are compared with the target side verb chunks while all the other chunks on the source side are compared with all the other chunks on the target side. Head words are extracted for each source chunk and the translated head words are actually compared on the target side taking into the consideration the synonymous target words. When the validation system returns positive, the source chunk is aligned with the identified original target chunk.

4 Tools and Resources used

A sentence-aligned English-Bengali parallel corpus containing 14,187 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 Phase II[3]". The Stanford Parser[4], Stanford NER, CRF chunker[5] and the Wordnet 3.0[6] have been used for identifying complex predicates in the source English side of the parallel corpus.

The sentences on the target side (Bengali) are parsed and POS-tagged by using the tools obtained from the consortium mode project "Development of Indian Language to Indian Language Machine Translation (IL-ILMT) System Phase II". NEs in Bengali are identified using the NER system of Ekbal and Bandyopadhyay (2008).

The effectiveness of the MWE-aligned and chunk 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).

5 Experiments and Evaluation Results

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 contains 13,176 sentences. In addition to the target side of the parallel corpus, a monolingual Bengali corpus containing 293,207 words from the tourism domain was used for the target language model. The experiments have been carried out 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 4 produce the optimum baseline result. The rest of the experiments have been carried out using these settings.

---

[3] The EILMT and ILILMT projects are funded by the Department of Information Technology (DIT), Ministry of Communications and Information Technology (MCIT), Government of India.

[4] http://nlp.stanford.edu/software/lex-parser.shtml

[5] http://crfchunker.sourceforge.net/

[6] http://wordnet.princeton.edu/

The system continues with the various preprocessing of the corpus. The hypothesis is that as more and more MWEs and chunks are identified and aligned properly, the system shows the improvement in the translation procedure. Table 1 shows the MWE statistics of the parallel training corpus. It is observed from Table 1 that NEs occur with high frequency in both sides compared to other types of MWEs. It suggests that prior alignment of the NEs and complex predicates plays a role in improving the system performance.

Training set	English		Bengali	
	T	U	T	U
CPs	4874	2289	14174	7154
reduplicated word	-	-	85	50
Noun-noun compound	892	711	489	300
Phrasal preposition	982	779	-	-
Phrasal verb	549	532	-	-
Total NE words	22931	8273	17107	9106

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

Single tokenization of NEs and MWEs of any length on both the sides followed by GIZA++ alignment has given a huge impetus to system performance (6.38 BLEU points absolute, 73% relative improvement over the baseline). In the source side, the system treats the phrasal prepositions, verb-object combinations and noun-noun compounds as a single token. In the target side, single tokenization of reduplicated phrases and noun-noun compounds has been done followed by alignments using the GIZA++ tool. From the observation of Table 2, during first iteration there are 81821 chunks are identified from the source corpus and 14534 has been aligned by the system. For iteration 2, there are 67287 source chunks are remaining to align. At the final iteration almost 65% of the source chunks have been aligned.

Training set	English		Bengali	
Iteration	T	U	T	U
1	81821	70321	65429	59627
2	67287	62575	50895	47139
final	32325	31409	15933	15654

Table 2. Chunk Statistics. (T - Total occurrence, U – Unique)

The system performance improves when the alignment list of NEs and complex predicates as well as sentence level aligned chunk are incorporated in the baseline best system. It achieves the BLEU score of 17.37 after the final iteration. This is the best result obtained so far with respect to the baseline system (8.63 BLEU points absolute, 98.74% relative improvement in Table 3). It may be observed from Table 3 that baseline Moses without any preprocessing of the dataset produces a BLEU score of 8.74.

Experiments		Exp	BLEU	NIST
Baseline		1	8.74	3.98
Best System (Alignment of NEs and Complex Predicates and Single Tokenization of various MWEs)		2	15.12	4.48
Baseline Best System + Chunk Alignment	Iteration 1	3	15.87	4.49
	Iteration 2	4	16.28	4.51
	Iteration 3	5	16.40	4.51
	Iteration 4	6	16.68	4.52
	Final Iteration†	**7**	**17.37**	**4.55**

Table 3. Evaluation results for different experimental setups. (The '†' marked systems produce statistically significant improvements on BLEU over the baseline system)

Intrinsic evaluation of the chunk alignment could not be performed as gold-standard word alignment was not available. Thus, 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). Bengali is a morphologically rich language and has relatively free phrase order. Proper evaluation of the English-Bengali

MT evaluation ideally requires multiple set of reference translations. Moreover, the training set was smaller in size.

6. Conclusions and Future work

A methodology has been presented in this paper to show how the simple yet effective preprocessing of various types of MWEs and alignment of NEs, complex predicates and chunks can boost the performance of PB-SMT system on an English—Bengali translation task. The best system yields 8.63 BLEU points improvement over the baseline, a 98.74% relative increase. A subset of the output from the best system has been compared with that of the baseline system, and the output of the best system almost always looks better in terms of either lexical choice or word ordering. It is observed that only 28.5% of the test set NEs appear in the training set, yet prior automatic alignment of the NEs complex predicates and chunk improves the translation quality. This suggests that not only the NE alignment quality in the phrase table but also the word alignment and phrase alignment quality improves significantly. At the same time, single-tokenization of MWEs makes the dataset sparser, but improves the quality of MT output to some extent. Data-driven approaches to MT, specifically for scarce-resource language pairs for which very little parallel texts are available, should benefit from these preprocessing methods. Data sparseness is perhaps the reason why single-tokenization of NEs and compound verbs, both individually and in collaboration, did not add significantly to the scores. However, a significantly large parallel corpus can take care of the data sparseness problem introduced by the single-tokenization of MWEs.

Acknowledgement

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

References

Agarwal, Aswini, Biswajit Ray, Monojit Choudhury, Sudeshna Sarkar and Anupam Basu. 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.(2004)

Baldwin, Timothy and Su Nam Kim 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 (2010)

Banerjee, Satanjeev, and Alon Lavie.. 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. (2005)

Carpuat, Marine, and Mona Diab. 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 (2010)

Chakraborty, Tanmoy and Sivaji Bandyopadhyay. 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. (2010)

Das, Dipankar, Santanu Pal, Tapabrata Mondal, Tanmoy Chakraborty, Sivaji Bandyopadhyay. 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.(2010)

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

Eck, Matthias, Stephan Vogel, and Alex Waibel. 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 (2004)

Ekbal, Asif, and Sivaji Bandyopadhyay. Voted NER system using appropriate unlabeled data. In proc. of the ACL-IJCNLP-2009 Named Entities Workshop (NEWS 2009), Suntec, Singapore, pp.202-210 (2009).

Huang, Young-Sook, Kyonghee Paik, Yutaka Sasaki, "Bilingual Knowledge Extraction Using Chunk Alignment", PACLIC 18, Tokiyo, pp. 127-138, (2004).

Kneser, Reinhard, and Hermann Ney. 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. (1995)

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 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. (2003)

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. 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. (2007)

Koehn, Philipp. 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. (2004)

Ma, Yanjun, Nicolas Stroppa, AndyWay. Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, ,Prague, Czech Republic, June 2007, pp. 304–311 (2007).

Moore, Robert C. Learning translations of named-entity phrases from parallel corpora. In Proc. of 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003), Budapest, Hungary; pp. 259-266. (2003)

Och, Franz J. 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. (2003)

Pal Santanu, Sudip Kumar Naskar, Pavel Pecina, Sivaji Bandyopadhyay and Andy Way. 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 (2010)

Pal, Santanu Tanmoy Chakraborty , Sivaji Bandyopadhyay, “Handling Multiword Expressions in Phrase-Based Statistical Machine Translation”, Machine Translation Summit XIII(2011),Xiamen, China, pp. 215-224 (2011)

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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 (2002)

Rayson, Paul, Dawn Archer, Scott Piao, and Tony McEnery. The UCREL Semantic Analysis System. In proc. Of LREC-04 Workshop: Beyond Named Entity Recognition Semantic Labeling for NLP Tasks, pages 7-12, Lisbon, Porugal (2004)

Ren, Zhixiang, Yajuan Lü, Jie Cao, Qun Liu, and Yun Huang. 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 (2009).

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. 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 (2003)

Wu, Hua Haifeng Wang, and Chengqing Zong. 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 (2008)

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

Zhou, Yu, chengquing Zong, Bo Xu, “Bilingual Chunk Aliment in Statistical Machine Translation”, IEEE International Conference on Systems, Man and Cybernetics, pp. 1401-1406, (2004)