A Hybrid Word Alignment Model for 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

This paper proposes a hybrid word alignment model for Phrase-Based Statistical Machine translation (PB-SMT). The proposed hybrid alignment model provides most informative alignment links which are offered by both unsupervised and semi-supervised word alignment models. Two unsupervised word alignment models (GIZA++ and Berkeley aligner) and a rule based aligner are combined together. The rule based aligner only aligns named entities (NEs) and chunks. The NEs are aligned through transliteration using a joint source-channel model. Chunks are aligned employing a bootstrapping approach by translating the source chunks into the target language using a baseline PB-SMT system and subsequently validating the target chunks using a fuzzy matching technique against the target corpus. All the experiments are carried out after single-tokenizing the multi-word NEs. Our best system provided significant improvements over the baseline as measured by BLEU.

1 Introduction

Word alignment is the backbone of PB-SMT system or any data driven approaches to Machine Translation (MT) and it has received a lot of attention in the area of statistical machine translation (SMT) (Brown et al., 1993; Och and Ney, 2003; Koehn et al., 2003). Word alignment is not an end task in itself and is usually used as an intermediate step in SMT. Word alignment is defined as the detection of corresponding alignment of words from parallel sentences that are translation of each other. Statistical machine translation usually suffers from many-to-many word links which existing statistical word alignment algorithms can not handle well.

The unsupervised word alignment models are based on IBM models 1–5 (Brown et al., 1993) and the HMM model (Ney and Vogel, 1996; Och and Ney, 2003). Models 3, 4 and 5 are based on fertility based models which are asymmetric. To improve alignment quality, the Berkeley Aligner is based on the symmetric property by intersecting alignments induced in each translation direction.

In the present work, we propose improvement of word alignment quality by combining three word alignment tables (i) GIZA++ alignment (ii) Berkeley Alignment and (iii) rule based alignment. Our objective is to perceive the effectiveness of the Hybrid model in word alignment by improving the quality of translation in the SMT system. In the present work, we have implemented a rule based alignment model by considering several types of chunks which are automatically extracted on the source side. Each individual source chunk is translated using a baseline PB-SMT system and validated with the target chunks on the target side. The validated source-target chunks are added in the rule based alignment table. Work has been carried out into three directions: (i) three alignment tables are combined together by taking their union; (ii) extra alignment pairs are added into the alignment table. This is a well-known practice in domain adaptation in SMT (Eck et al., 2004; Wu et al., 2008); (iii) the alignment table is updated through semi-supervised alignment technique.

Proceedings of the Second Workshop on Hybrid Approaches to Translation, pages 94–101,
Sofia, Bulgaria, August 8, 2013.

The remainder of the paper is organized as follows. Section 2 discusses related work. The proposed hybrid word alignment model is described in Section 3. Section 4 presents 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 Works

Zhou et al. (2004) proposed a multi lingual filtering algorithm that generates bilingual chunk alignment from Chinese-English parallel corpus. The algorithm has three steps, first, from the parallel corpus; the most frequent bilingual chunks are extracted. Secondly, the participating chunks for alignments are combined into a cluster 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). Pal et, al. (2012) proposed a bootstrapping method for chunk alignment; they used an SMT based model for chunk translation and then aligned the source-target chunk pairs after validating the translated chunk. 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. A Maximum Entropy model based approach for English—Chinese NE alignment which significantly outperforms IBM Model4 and HMM has been proposed by Feng et al. (2004). They considered 4 features: translation score, transliteration score, source NE and target NE's co-occurrence score and the distortion score for distinguishing identical NEs in the same sentence. Moore (2003) presented an approach where capitalization cues have been used for identifying NEs on the English side. Statistical techniques are applied to decide which portion of the target language corresponds to the specified English NE, for simultaneous NE identification and translation.

To improve the learning process of unlabeled data using labeled data (Chapelle et al., 2006), the semi-supervised learning method is the most useful learning technique. Semi-supervised learning is a broader area of Machine Learning. Researchers have begun to explore semi-supervised word alignment models that use both labeled and unlabeled data. Fraser and Marcu (2006) proposed a semi-supervised training algorithm. The weighting parameters are learned from discriminative error training on labeled data, and the parameters are estimated by maximum-likelihood EM training on unlabeled data. They have also used a log-linear model which is trained on the available labeled data to improve performance. Interpolating human alignments with automatic alignments has been proposed by Callison-Burch et al. (2004), where the alignments of higher quality have gained much higher weight than the lower-quality alignments. Wu et al. (2006) have developed two separate models of standard EM algorithm which learn separately from both labeled and unlabeled data. Two models are then interpolated as a learner in the semi-supervised Ada-Boost algorithm to improve word alignment. Ambati et al. (2010) proposed active learning query strategies to identify highly uncertain or most informative alignment links under an unsupervised word alignment model.

Intuitively, multiword NEs 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 multiword NE are marked and aligned as parts of consecutive phrases, since PB-SMT (or any other approaches to SMT) does not generally treat multiword NEs as special tokens. This is the motivations behind considering NEs for special treatment in this work by converting into single tokens that makes sure that PB-SMT also treats them as a whole

Another problem with SMT systems is the erroneous word alignment. Sometimes some words are not translated in the SMT output sentence because of the mapping to NULL token or erroneous mapping during word alignment. Verb phrase translation also creates major problems. 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.

The first objective of the present work is to see how single tokenization and alignment of NEs on both the sides affects the overall MT quality. The second objective is to see whether Hybrid word alignment model of both unsupervised and semi-supervised techniques enhance the quality of translation in the SMT system rather than the single tokenized NE level parallel corpus applied to the hybrid model.

We carried out the experiments on English—Bengali translation task. Bengali shows high morphological richness at lexical level. Lan-

guage resources in Bengali are not widely available.

3 Hybrid Word Alignment Model

The hybrid word alignment model is described as the combination of three word alignment models as follows:

3.1 Word Alignment Using GIZA++

GIZA++ (Och and Ney, 2003) is a statistical word alignment tool which incorporates all the IBM 1-5 models. GIZA++ facilitates fast development of statistical machine translation (SMT) systems. In case of low-resource language pairs the quality of word alignments is typically quite low and it also deviates from the independence assumptions made by the generative models. Although huge amount of parallel data enables the model parameters to acquire better estimation, a large number of language pairs still lacks from the unavailability of sizeable amount of parallel data. GIZA++ has some draw-backs. It allows at most one source word to be aligned with each foreign word. To resolve this issue, some techniques have already been applied such as: the parallel corpus is aligned bidirectionally; then the two alignment tables are reconciled using different heuristics e.g., intersection, union, and most recently grow-diagonal-final and grow-diagonal-final-and heuristics have been applied. In spite of these heuristics, the word alignment quality for low-resource language pairs is still low and calls for further improvement. We describe our approach of improving word alignment quality in the following three subsections.

3.2 Word Alignment Using Berkley Aligner

The recent advancements in word alignment is implemented in Berkeley Aligner (Liang et al., 2006) which allows both unsupervised and supervised approach to align word from parallel corpus. We initially train the parallel corpus using unsupervised technique. We make a few manual corrections to the alignment table produced by the unsupervised aligner. Then we apply this corrected alignment table as gold standard training data for the supervised aligner. The Berkeley aligner is an extension of the Cross Expectation Maximization word aligner. Berkeley aligner is a very useful word aligner because it allows for supervised training, enabling us to derive knowledge from already aligned parallel corpus or we can use the same corpus by updating the alignments using some rule based methods. Our approach deals with the latter case. The supervised technique of Berkeley aligner helps us to align those words which could not be aligned by rule based word aligner.

3.3 Rule Based Word Alignment

The proposed Rule based aligner aligns Named Entities (NEs) and chunks. For NE alignment, we first identify NEs from the source side (i.e. English) using Stanford NER. The NEs on the target side (i.e. Bengali) are identified using a method described in (Ekbal and Bandyopadhyay, 2009). The accuracy of the Bengali Named Entity recognizers (NER) is much poorer compared to that of English NER due to several reasons: (i) there is no capitalization cue for NEs in Bengali; (ii) most of the common nouns in Bengali are frequently used as proper nouns; (iii) suffixes (case markers, plural markers, emphasizers, specifiers) get attached to proper names as well in Bengali. Bengali shallow parser[1] has been used to improve the performance of NE identification by considering proper names as NE. Therefore, NER and shallow parser are jointly employed to detect NEs from the Bengali sentences. The source NEs are then transliterated using a modified joint source-channel model (Ekbal et al., 2006) and aligned to their target side equivalents following the approach of Pal et al. (2010). The target side equivalents NEs are transformed into canonical form after omitting their '*matras*'. Similarly Bengali NEs are also transformed into canonical forms as Bengali NEs may differ in their choice of *matras* (vowel modifiers). The transliterated NEs are then matched with the corresponding parallel target NEs and finally we align the NEs if match is found.

After identification of multiword NEs on both sides, we pre-processed the corpus by replacing space with the underscore character ('_'). We have used underscore ('_') instead of hyphen ('-') since there already exists some hyphenated words in the corpus. The use of the underscore ('_') character also facilitates to de-tokenize the single-tokenized NEs after decoding.

For chunk alignment, the source sentences of the parallel corpus are parsed using Stanford POS tagger. The chunks of the sentences are extracted using CRF chunker[2]. The chunker detects the boundaries of noun, verb, adjective, adverb

[1] http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

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

and prepositional chunks from the sentences. In case of prepositional phrase chunks, we have taken a special attention: we have expanded the prepositional phrase chunk by examining a single noun chunk followed by a preposition or a series of noun chunks separated by conjunctions such as *'comma', 'and'* etc. For each individual chunk, the head word is identified. Similarly target side sentences are parsed using a shallow parser. The individual target side Bengali chunks are extracted from the parsed sentences. The head words for all individual chunks on the target side are also marked. If the translated head word of a source chunk matches with the headword of a target chunk then we hypothesize that these two chunks are translations of each other.

The extracted source chunks are translated using a baseline SMT model trained on the same corpus. The translated chunks are validated against the target chunks found in the corresponding target sentence. During the validation process, if any match is found between the translated chunk and a 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 and is considered in the next iterations.

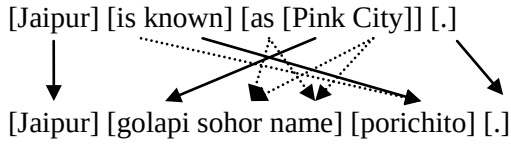

Figure 1.a: Rule based alignments

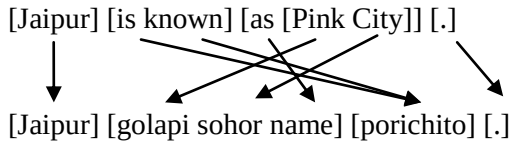

Figure 1.b: Gold standard alignments

Figure 1: Establishing alignments through Rule based methods.

The extracted chunks on the source side may not have a one to one correspondence with the target side chunks. The alignment validation process is focused on the proper identification of the head words and not between the translated source chunk and target chunk. The matching process has been carried out using a fuzzy matching technique. If both sides contain only one chunk after aligning the remaining chunks then the alignment is trivial. After aligning the individual chunks, we also establish word alignments between the matching words in those aligned chunks. Thus we get a sentence level source-target word alignment table.

Figure 1 shows how word alignments are established between a source-target sentence pair using the rule based method. Figure 1.a shows the alignments obtained through rule based method. The solid links are established through transliteration (for NEs) and translation. The dotted arrows are also probable candidates for intra-chunk word alignments; however they are not considered in the present work. Figure 1.b shows the gold standard alignments for this sentence pair.

3.4 Hybrid Word alignment Model

The hybrid word alignment method combines three different kinds of word alignments − Giza++ word alignment with grow-diag-final-and (GDFA) heuristic, Berkeley aligner and rule based aligner. We have followed two different strategies to combine the three different word alignment tables.

Union
In the union method all the alignment tables are united together and duplicate entries are removed.

ADD additional Alignments
In this method we consider either of the alignments generated by GIZA++ GDFA (A1) or Berkeley aligner (A2) as the standard alignment as the rule based aligner fails to align all words in the parallel sentences. From the three set of alignments A1, A2 and A3, we propose an alignment combination method as described in algorithm 1.

ALGORITHM: 1

Step 1: Choose either A1 or A2 as the standard alignment (SA).
Step 2: Correct the alignments in SA using the alignment table of A3.
Step 3: if A2 is considered as SA then find additional alignment from A1 and A3 using intersection method (A1∩A3) otherwise find additional alignment from A2 and A3 (using A2∩A3).
Step 4: Add additional entries with SA.

3.5 Berkeley Semi-supervised Alignment

The correctness of the alignments is verified by manually checking the performance of the various alignment system. We start with the combined alignment table which is produced by Algorithm 1. Iinitially, we take a subset of the alignments by manually inspecting from the combined alignment table. Then we train the Barkley supervised aligner with this labeled data. A subset of the unlabeled data from the combined alignment table is tested with the supervised model. The output is then added as additional labeled training data for the supervised training method for the next iteration. Using this bootstrapping approach, the amount of labeled training data for the supervised aligner is gradually increased. The process is continued until there are no more unlabelled training data. In this way we tune the whole alignment table for the entire parallel corpus. The process is carried out in a semi-supervised manner.

4 Tools and resources Used

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 - Phase II"[3]. The Stanford Parser[4] and CRF chunker[5] have been used for identifying chunks and Stanford NER has been used to identify named entities in the source side of the parallel corpus.

The target side (Bengali) sentences are parsed by using the tools obtained from the consortium mode project "Development of Indian Language to Indian Language Machine Translation (IL-ILMT) System - Phase II[6]".

The effectiveness of the present work has been tested by using the standard log-linear PB-SMT model as our baseline system: phrase-extraction heuristics described in (Koehn et al., 2003), , MERT (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 Results

We have randomly selected 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 building 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 produced the optimum baseline result. We carried out the rest of the experiments using these settings.

We experimented with the system over various combinations of word alignment models. Our hypothesis focuses mainly on the theme that proper alignment of words will result in improvement of the system performance in terms of translation quality.

141,821 chunks were identified from the source corpus, of which 96,438 (68%) chunks were aligned by the system. 39,931 and 28,107 NEs were identified from the source and target sides of the parallel corpus respectively, of which 22,273 NEs are unique in English and 22,010 NEs in Bengali. A total of 14,023 NEs have been aligned through transliteration.

The experiments have been carried out with various experimental settings: (i) single tokenization of NEs on both sides of the parallel corpus, (ii) using Berkeley Aligner with unsupervised training, (iii) union of the three alignment models: rule based, GIZA++ with GDFA and Berkeley Alignment, (iv) hybridization of the three alignment models and (v) supervised Berkeley Aligner. Eextrinsic evaluation was carried out on the MT quality using BLEU (Papineni et al., 2002) and NIST (Doddington, 2002).

---

[3] The EILMT project is funded by the Department of Electronics and Information Technology (DEITY), 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] 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.

Experiment	Exp no.	BLEU	NIST
Baseline system using GIZA++ with GDFA	1	10.92	4.13
PB-SMT system using Berkeley Aligner	2	11.42	4.16
Union of all Alignments	3	11.12	4.14
PB-SMT System with Hybrid Alignment by considering (a) GIZA++ as the standard alignment) (b) Berkeley alignment as the standard alignment)	4a†	15.38	4.30
	4b†	15.92	4.36
Single tokenized NE + Exp 1	5	11.68	4.17
Single tokenized NE + Exp 2	6	11.82	4.19
Single tokenized NE + (a) Exp 4a (b) Exp 4b	7a†	16.58	4.45
	7b†	17.12	4.49
PB-SMT System with semi-supervised Berkeley Aligner + Single tokenized NE	8†	**20.87**	**4.71**

Table: 1 Evaluation results for different experimental setups. (The '†' marked systems produce statistically significant improvements on BLEU over the baseline system)

The baseline system (Exp 1) is the state-of-art PB-SMT system where GIZA++ with grow-diag-final-and has been used as the word alignment model. Experiment 2 provides better results than experiment 1 which signifies that Berkeley Aligner performs better than GIZA++ for the English-Bengali translation task. The union of all thee alignments (Exp 3) provides better scores than the baseline; however it cannot beat the results obtained with the Berkeley Aligner alone.

Hybrid alignment model with GIZA++ as the standard alignment (Exp 4a) produces statistically significant improvements over the baseline. Similarly the use of Berkeley Aligner as the standard alignment for hybrid alignment model (Exp 4b) also results in statistically significant improvements over Exp 2. These two experiments (Exp 4a and 4b) demonstrate the effectiveness of the hybrid alignment model. It is to be noticed that hybrid alignment model works better with the Berkeley Aligner than with GIZA++.

Single-tokenization of the NEs (Exp 5, 6, 7a and 7b) improves the system performance to some extent over the corresponding experiments without single-tokenization (Exp 1, 2, 4a and 4b); however, these improvements are not statistically significant. The Berkeley semi-supervised alignment method using a bootstrapping approach together with single-tokenization of NEs provided the overall best performance in terms of both BLEU and NIST and the corresponding improvement is statistically significant on BLEU over rest of the experiments.

6 Conclusion and Future Work

The paper proposes a hybrid word alignment model for PB-SMT. The paper also shows how effective pre-processing of NEs in the parallel corpus and direct incorporation of their alignment in the word alignment model can improve SMT system performance. In data driven approaches to MT, specifically for scarce resource data, this approach can help to upgrade the state-of-art machine translation quality as well as the word alignment quality. . The hybrid model with the use of the semi-supervised technique of the Berkeley word aligner in a bootstrapping manner, together with single tokenization of NEs, provides substantial improvements (9.95 BLEU points absolute, 91.1% relative) over the baseline. On manual inspection of the output we found that our best system provides more accu-

rate lexical choice as well as better word ordering than the baseline system.

As future work we would like to explore how to get the best out of multiple word alignments. Furthermore, integrating the knowledge about multi-word expressions into the word alignment models is another future direction for this work.

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

Alexander Fraser and Daniel Marcu. 2006. Semi-supervised training for statistical word alignment. In ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-2006), Morristown, NJ, USA. pages 769–776.

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.

Chris Callison-Burch, David Talbot, and Miles Osborne. 2004. Statistical machine translation with word- and sentence-aligned parallel corpora. In ACL 2004, page 175, Morristown, NJ, USA. Association for Computational Linguistics.

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 cooccurrence statistics. In Proceedings of the Second International Conference on Human Language Technology Research (HLT-2002), San Diego, CA, pp. 128-132.

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.

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

Feng, Donghui, Yajuan Lv, and Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), Barcelona, Spain, pp. 372-379.

Feng, Donghui, Yajuan Lv, and Ming Zhou. 2004. A new approach for English-Chinese named entity alignment. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), Barcelona, Spain, pp. 372-379.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. Computational Linguistics, pages 19–51.

Huang, Fei, Stephan Vogel, and Alex Waibel. 2003. Automatic extraction of named entity translingual equivalence based on multi-feature cost minimization. In Proceedings of the ACL-2003 Workshop on Multilingual and Mixed-language Named Entity Recognition, 2003, Sapporo, Japan, pp. 9-16.

HuaWu, HaifengWang, and Zhanyi Liu. 2006. Boosting statistical word alignment using labeled and unlabeled data. In Proceedings of the COLING/ACL on Main conference poster sessions, pages 913–920, Morristown, NJ, USA. Association for Computational Linguistics.

Kneser, Reinhard, and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In Proceedings 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 Proceedings 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 Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL 2007): Proceedings of demo and poster sessions, Prague, Czech Republic, pp. 177-180.

Koehn, Philipp. 2004. Statistical significance tests for machine translation evaluation. In EMNLP-2004:

Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 25-26 July 2004, Barcelona, Spain, pp 388-395.

O. Chapelle, B. Sch¨olkopf, and A. Zien, editors. 2006. Semi-Supervised Learning. MIT Press, Cambridge, MA.

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

Pal, Santanu, Sivaji Bandyopadhyay. 2012, "Bootstrapping Chunk Alignment in Phrase-Based Statistical Machine Translation", Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), EACL-2012, Avignon, France, pp. 93-100 .

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 Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), Philadelphia, PA, pp. 311-318.

Percy Liang, Ben Taskar, Dan Klein. 2006. 6th Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL-2006, Pages 104-111

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

Vamshi Ambati, Stephan Vogel, Jaime Carbonell. 2010, 10th Proceedings of the NAACL HLT 2010 Workshop on Active Learning for Natural Language Processing (ALNLP-2010), Pages 10-17.

Vogel, Stephan, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In Proc. of the 16th International Conference on Computational Linguistics (COLING 1996), Copenhagen, pp. 836-841.

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.

X. Zhu. 2005. Semi-Supervised Learning Literature Survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison. http://www.cs.wisc.edu/_jerryzhu/pub/ssl_survey.pdf.