SEMANTICALLY DRIVEN INVERSION TRANSDUCTION GRAMMAR INDUCTION FOR EARLY STAGE TRAINING OF SPOKEN LANGUAGE TRANSLATION

Meriem Beloucif and Dekai Wu

HKUST Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
{mbeloucif | dekai@cs.ust.hk}

***Index Terms*—** spoken language translation, semantic SMT, semantic MT evaluation, semantic role labeling, low resource SMT

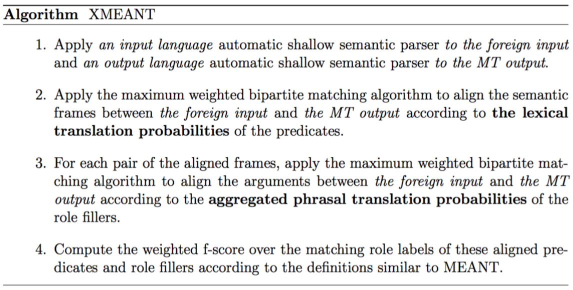
ABSTRACT

We propose an approach in which we inject a crosslingual semantic frame based objective function directly into inversion transduction grammar (ITG) induction in order to semantically train spoken language translation systems. This approach represents a follow-up of our recent work on improving machine translation quality by tuning loglinear mixture weights using a semantic frame based objective function in the late, final stage of statistical machine translation training. In contrast, our new approach injects a semantic frame based objective function back into *earlier* stages of the training pipeline, during the actual learning of the translation model, biasing learning toward semantically more accurate alignments. Our work is motivated by the fact that ITG alignments have empirically been shown to fully cover crosslingual semantic frame alternations. We show that injecting a crosslingual semantic based objective function for driving ITG induction further sharpens the ITG constraints, leading to better performance than either the conventional ITG or the traditional GIZA++ based approaches.

1. INTRODUCTION

In this paper, we show that injecting a crosslingual semantic based objective function at a very early stage of training spoken language translation systems improves the translation quality compared to using conventional approaches. Research has previously shown that including a semantic based objective function much later in the training pipeline by tuning against semantic based metrics, MEANT [1], improves the translation adequacy [2, 3, 4, 5]. We show that injecting a semantic based objective function much earlier in the training produces more semantically correct translations. Our approach is highly motivated by the idea behind our recently developed crosslingual evaluation metric, XMEANT [6]. We apply an XMEANT based crosslingual semantic frame alignment method for constraining inversion transduction grammars (ITGs). We show that this way of inducing ITGs helps to learn more semantically valid alignments compared to both conventional ITGs and the traditional GIZA++ alignments, leading to better translations. Our approach is motivated by the fact that XMEANT has been shown to correlate better wit human adequacy judgement than most of the commonly used metrics [6]. Furthermore, ITG alignments have previously been empirically shown to almost fully cover crosslingual semantic frame alternations, even though they rule out the majority of incorrect alignments [7]. We show that using XMEANT-like semantic frame matching for inducing ITGs not only helps to further narrow down inversion transduction grammar constraints, but also avoids losing relevant portions of the search space, leading to more semantically driven word alignments. We also show that a semantic based learning can help to improve the translation quality for low resource languages in comparison to existing learning methods by deliberately training our approach using a relatively small dataset. We adopt DARPA's approach in the LORELEI dry run evaluation, simulating low resource conditions in a Chinese-English translation learning task (despite the fact that Chinese is not a low resource language) by deliberately restricting the parallel training data to a small dataset, namely the *International Workshop on Spoken Language Translation (IWSLT07)*, and show that our method outperforms the traditional alignment methods for spoken data.

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under LORELEI contract HR0011-15-C-0114, BOLT contracts HR0011-12-C-0014 and HR0011-12-C-0016, and GALE contracts HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the Horizon 2020 grant agreement 645452 (QT21) and FP7 grant agreement 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF16210714, GRF16214315, GRF620811 and GRF621008. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA, the EU, or RGC.



Algorithm XMEANT

1. Apply *an input language* automatic shallow semantic parser *to the foreign input* and *an output language* automatic shallow semantic parser *to the MT output.*
2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between *the foreign input* and *the MT output* according to **the lexical translation probabilities** of the predicates.
3. For each pair of the aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between *the foreign input* and *the MT output* according to the **aggregated phrasal translation probabilities** of the role fillers.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers according to the definitions similar to MEANT.

Fig. 1. Algorithm of XMEANT

2. RELATED WORK

2.1. Word alignment

Word alignment is considered to be an important step in training MT systems, since it helps to learn the correlations between the input and the output languages. Unfortunately, conventional alignments are generally based on training IBM models [8], which are known to produce weak word alignment since they allow unstructured movement of words. Then take the intersection of alignments in both directions to produce the final alignment.

A hidden Markov model (HMM) based alignment was proposed by Vogel *et al.* [9], but similarly to IBM models, the objective function uses surface based alignment rather than a more structure based alignment. No constraints are used while training, allowing many random word-to-word permutations. Such an alignment generally hurts translation accuracy and adequacy.

For producing word alignments via unsupervised training of inversion transduction grammars [10], a method with improved efficiency has been developed in this work starting with Saers *et al.* [11]. This method tackles the issue that exhaustive biparsing and training using ITGs requires $O(n^6)$ time which, though feasible, is slow; instead, an improved method runs in $O(n^3)$ time [12].

In this work, we use BITGs or bracketing transduction grammars [11], which only use one single nonterminal category and surprisingly achieve good results by outperforming the conventional GIZA++ alignments [13]. It has been shown that ITG constraints allow higher flexibility in word ordering for longer sentences than the conventional IBM model, and that applying ITG constraints for word alignment leads to learning a significantly better alignment than the constraints used in conventional IBM models for both German-English and French-English language pairs [14]. In a version of ITGs proposed by Zhang and Gildea [15], rule probabilities are lexicalized throughout the biparse tree for efficient training, which helps to align sentences up to 15 words.

Some of the previous work on word alignment used morphological and syntactic features [16]. Some log linear models have been proposed to incorporate these features [17]. The problem with these approaches is that they require language specific knowledge and that they always work better on more morphologically rich languages.

A few studies that approximately integrate semantic knowledge in computing word alignment have been proposed [18], [19]. However, the former needs to have a prior word alignment learned on lexical words. The authors of the latter model proposed a semantic oriented word alignment. However, they need to extract word similarity from the monolingual data first then produce alignment using word similarities.

2.2. XMEANT: crosslingual MEANT

Our method is fully consistent with the principle adopted by the MEANT family of metrics, in which a good output translation is one where the core semantic of the input sentence is preserved, as captured by the basic event structure *who did what to whom, for whom, when, where, how and why* [20]. MEANT is a weighted f-score over the matched semantic role labels of automatically aligned semantic frames and role fillers [1, 21, 22]. It evaluates the degree of goodness of the MT output sentence against the provided reference translations, and produces a score that measures the degree of similarity between their semantic frame structures. Our new approach is encouraged by the fact that many previous studies have empirically shown that integrating semantic role labeling into the training pipeline by tuning against MEANT improves the translation adequacy [2, 3, 4, 5].

Unlike n-gram or edit-distance based metrics, the MEANT family of metrics adopt the principle that a good translation is one in which humans can successfully understand the general meaning of the input sentence as captured by the basic event structure defined in [20]. Recent works have shown that the semantic frame based metric, MEANT, correlates better with human adequacy judgment than the most common evaluation metrics such as BLEU [23], NIST [24], METEOR [25], CDER [26], WER [27], and TER [28]. It has been shown that including semantic role labeling in the training pipeline by tuning against a semantic frame objective function such as the semantic evaluation metric MEANT significantly improves the quality of the MT output. Previous work has shown [29] that injecting a crosslingual objective function into the training pipeline helps to improve the quality of the word alignment. We argue in this paper that, incorporating monolingual semantic information while training SMT systems can help to learn more semantically correct bilingual correlations for low resource languages, as in the DARPA LORELEI program.

The crosslingual XMEANT metric [6] has been shown to correlate even better with human adequacy judgments than MEANT. Unlike MEANT, which needs the expensive man made references for the MT evaluation, XMEANT uses the

$$
\begin{aligned}
\mathbf{e}_{i,\text{pred}} &\equiv \text{the output side of the pred of aligned frame } i \\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the input side of the pred of aligned frame } i \\
\mathbf{e}_{i,j} &\equiv \text{the output side of the ARG } j \text{ of aligned frame } i \\
\mathbf{f}_{i,j} &\equiv \text{the input side of the ARG } j \text{ of aligned frame } i \\
p(e,f) &= \sqrt{t\left(e|f\right)t\left(f|e\right)} \\
\text{prec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{e\in\mathbf{e}} \max_{f\in\mathbf{f}} p(e,f)}{|\mathbf{e}|} \\
\text{rec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{f\in\mathbf{f}} \max_{e\in\mathbf{e}} p(e,f)}{|\mathbf{f}|} \\
s_{i,\text{pred}} &= \frac{2\cdot\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}\cdot\text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}+\text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}} \\
s_{i,j} &= \frac{2\cdot\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}\cdot\text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}+\text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}
\end{aligned}
$$

Fig. 2. Crosslingual phrasal similarity in XMEANT

foreign input to evaluate the MT translation output. MEANT measures lexical similarity using a monolingual context vector model, whereas XMEANT substitutes simple crosslingual lexical translation probabilities. Figure 1 describes the XMEANT algorithm. XMEANT uses MEANT's f-score based method for aggregating lexical translation probabilities within the semantic role filler phrases. Each token of the role fillers in the output/input string is aligned to the token of the role fillers in the input/output string that has the maximum lexical translation probability. The crosslingual phrasal similarities are computed as shown in Figure 2.

Our approach uses the XMEANT method of matching semantic predicates and role labels between the input and the output, and uses this crucial information for inducing inversion transduction grammars. In this paper we show that by using this semantic objective function at an early stage of training the statistical machine translation (SMT) system, not only are we able to learn more semantic correlations between the two languages, but also that this holds even under low resource conditions limited to small amounts of parallel data, as in the DARPA LORELEI program.

3. DRIVING ITG INDUCTION USING XMEANT-LIKE SIMILARITY FUNCTION

In this paper, we propose a model that injects an XMEANT-like semantic frame based objective function into early stage SMT training, thereby biasing bracketing inversion transduction grammar induction (BITG) towards preferring more semantically valid bilingual constituents, that best fits XMEANT's crosslingual semantic frames. Giving the structural differences between the monolingual semantic parsers and the bilingual BITG parses, XMEANT penalizes BITG biconstituents that violate the crosslingually aligned semantic frames.

A penalty is paid whenever the BITG biparser wants to introduce a biconstituent crosses a semantic constituent (the predicat or one of its role fillers), making the biconstituent not fitting into XMEANT's alignment. In this way, a penalty is paid for biconstituents completely covering semantic biconstituents that are completely covered by semantic constituents. To allow for some degree of freedom, we allow for two penalty levels, one for crossing an input language semantic constituent, and one for crossing an output language semantic constituent. All these hyperparameters are set manually for now.

The semantic trees are not necessarily consistent with the syntactic trees, since the semantic roles and their fillers in a sentence sometimes span across multiple syntactic units. On the other hand, BITG trees are defined to be projective, thus applying even a single monolingual semantic parse would rule out *all* possible BITG trees, and *all* possible alignments for that sentence pair. As the lexical relation is what defines the word alignment, which is what we are interested in, XMEANT penalizes any constraints that violate XMEANT's semantic frame alignment. In practice, the automatic semantic shallow parses are fairly noisy, which is a reason to soften them.

4. EXPERIMENTAL SETUP

4.1. SMT pipeline

We compare the performance of our proposed XMEANT-driven alignments to the conventional ITG alignment and to the traditional GIZA++ baseline with grow-diag-final-and to harmonize both alignment directions. We also perform a grid search over the hyperparameters in our proposed model to find the optimal settings.

Our ITG baseline is a token-based BITG system. We initialize it with uniform structural probabilities, setting aside half of the probability mass for lexical rules. This probability mass is distributed among the lexical rules according to co-occurrence counts from the training data, assuming each sentence to contain one empty token to account for singletons. These initial probabilities are refined with 10 iterations of expectation maximization where the expectation step is calculated using beam pruned parsing [13] with a beam width of 100. On the last iteration, we extract the alignments imposed by the Viterbi parses as the word alignments output by the system.

Compared to the ITG baseline discussed above, our new model rewards any biconstituent that falls into an XMEANT semantic frame alignment, as discussed in Section 3. The shallow semantic parses of the training data were produced using ASSERT [20] and C-ASSERT [30] for English and Chinese respectively. The hyperparameters were only used during the training to set the probabilities of the grammar, not

Table 1. Translation quality of the three alignment methods used in Chinese-English MT systems using IWSLT 2007, trained using Moses hierarchical.

System	MEANT	BLEU	METEOR	TER	WER	PER	CDER
Giza alignment	49.94	23.02	4.14	59.95	60.52	55.58	59.14
ITG alignment	50.57	21.82	**4.32**	**57.86**	**58.68**	53.90	57.38
XMEANT-driven	**50.92**	**24.70**	4.27	58.44	59.01	**53.85**	57.58

Table 2. Translation quality of the three alignment methods used in Chinese-English MT systems using IWSLT 2007, trained using Moses phrase based.

System	MEANT	BLEU	METEOR	TER	WER	PER	CDER
Giza alignment	47.65	18.59	3.70	63.01	63.83	57.37	62.02
ITG alignment	48.36	18.44	**4.02**	**61.09**	**62.63**	**54.96**	**60.54**
XMEANT-driven	**48.56**	**20.35**	4.02	61.17	62.77	55.42	60.46

when extracting the Viterbi parses and the corresponding word alignments.

For our expermental setup, we purposely use a relatively small corpus to simulate low resource language scenario. We show that including a semantic based objective function during the actual learning of the SMT model helps better learning bilingual correlations, without relying on heavy memorization from expensive huge parallel corpora. Although Chinese is not a low resource language, we adopted the DARPA LORELEI program's approach in its dry run evaluation, by purposely simulating low resource conditions, in the present case by using a relatively small corpus (IWSLT07). The training set contains 39,953 sentences. The training set, development set, and test set were the same for all systems in order to keep the experiments comparable.

We tested the different alignments described above by using the standard MOSES toolkit [31], and a 6-gram language model learned with the SRI language model toolkit [32] to train our model. We tested our approach with both MOSES hierarchical and MOSES phrase based. For tuning, we used ZMERT [33] the standard implementation of minimum error rate training, or MERT [34].

5. RESULTS

Results show that our proposed model outperforms the conventional BITG based model and the traditional GIZA++ with GDFA as a heuristic, we tested the performance of of our proposed model with two baselines: MOSES phrase based and MOSES hierarchical. We evaluated our MT output using the semantic metric MEANT [1] and also surface based metrics such as BLEU [23], METEOR [25], CDER [26], WER [27], and TER [28].

We note from the results that the MEANT score for ITG with semantic constraints is slightly better than the conventional ITG model. We believe that a better shallow semantic parser would yield a better system. Our results show that we should be more focused on including semantic information while training SMT system rather than just tuning against a semantic objective function. Both ITG based systems give a comparable result which is still very high in comparison to GIZA++ alignment in term of edit distance metrics and MEANT score. Tables 1 and 2 show the interesting improvement in terms of BLEU and MEANT scores for our proposed XMEANT-driven aligned system in comparison to conventional BITG alignment and GIZA++ alignment for both Moses baselines. Both BLEU and MEANT scores for our new proposed alignment are considerably higher than the BLEU and MEANT scores for the conventional BITG and the traditional GIZA++ based systems.

Figure 4 shows an interesting example extracted from our translated data and compared to the translations obtained by other systems. We note from these examples that the more structured ITG based models give a more accurate output compared to the heuristic based GIZA++ alignment. Example 1 shows an interesting example in which the XMEANT-driven system learns a more accurate translation of the input sentence, whereas the GIZA++ fails completely to capture the basic semantics of the input. The ITG system on the other hand, correctly gets the global meaning of the input but fails to use the right wording (has come off). Example 2 shows an example where learning the right semantic structure can not only produce better adequacy, but also leads to a better fluency for low resource languages. We emphasize here, that both GIZA++ and ITG models fail to capture the right translation due to insufficient training data. The semantic frame based objective function that we used shows that by capturing the right structure while learning the alignment, we can produce better translations even when using a very small data set. Example 3 is also interesting in the sense that, having no context, both ITG and XMEANT output can be considered as valid translations. This shows again, that semantic based heuristics are needed for more disambiguation, on the other hand, GIZA++ based alignment fails to completely capture any meaning once again.

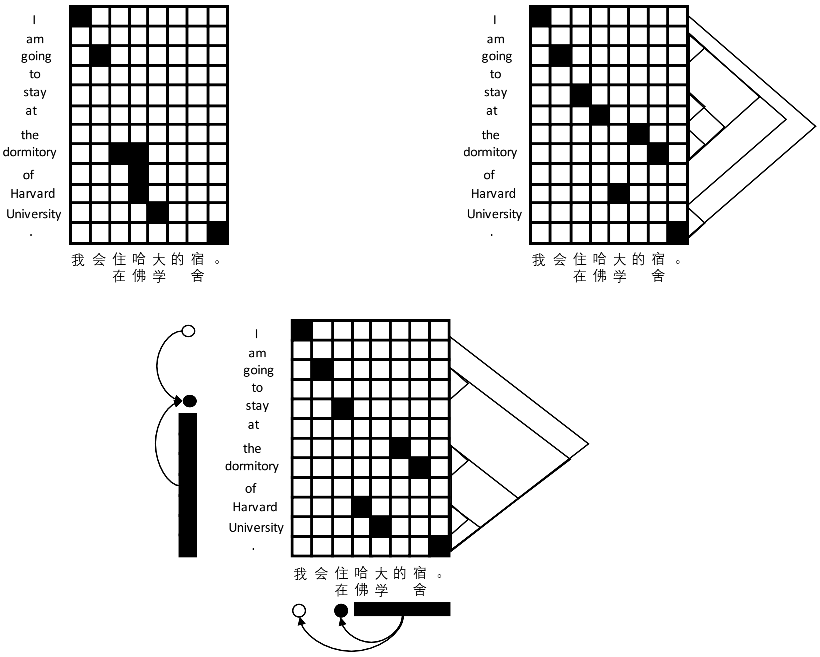
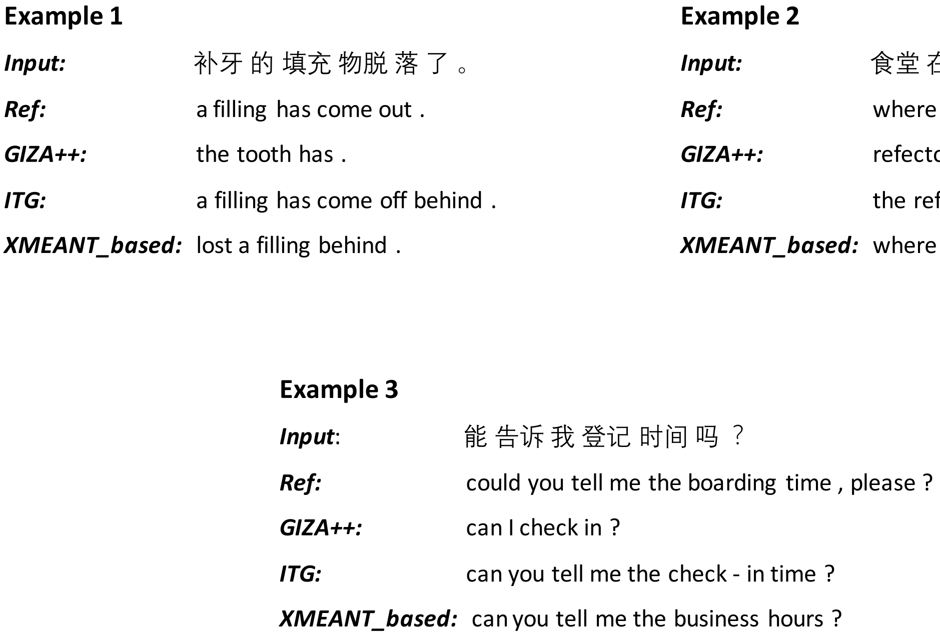


Fig. 3. An alignment of bi-sentences produced by both GIZA++ (left) and ITG based alignments (right) at the top of the picture, and the XMEANT-driven constrained ITG alignment at the bottom



Example 1

Input: 补牙 的 填充 物脱 落 了 。

Ref: a filling has come out .

GIZA++: the tooth has .

ITG: a filling has come off behind .

XMEANT_based: lost a filling behind .

Example 2

Input: 食堂 在 哪里 ？

Ref: where 's the dining room ?

GIZA++: refectory then where ?

ITG: the refectory where ?

XMEANT_based: where 's the refectory ?

Example 3

Input: 能 告诉 我 登记 时间 吗 ？

Ref: could you tell me the boarding time , please ?

GIZA++: can I check in ?

ITG: can you tell me the check - in time ?

XMEANT_based: can you tell me the business hours ?

Fig. 4. Interesting examples comparing the output of the three compared systems

Figure 3 represents the alignment obtained after running GIZA++, the ITG based system, and our new system baseline respectively. We observe that both GIZA++ and ITG alignments fail to align different crucial parts of the parallel sentences. The XMEANT-driven alignment gives a very good alignment based on the semantic structure of both semantic parsers. We see that it only fails while trying to align the to "的", which can be explained by the fact that, from either English-to-Chinese or Chinese-to-English, the word the or the character "的" will be translated to NULL. There are cases where "的" gets translated to other similar non-function-words such as 's or quotation marks, but we can consider these to detract relatively little from the general understandability of the translation.

6. CONCLUSION

In this paper we showed that injecting a semantic frame based objective function at a relatively early stage in the training of spoken language translation helps to improve the quality of the translation. We have presented an approach to semantically drive the learning of spoken language translation models, by constraining ITG with XMEANT like alignments. We have also demonstrated that using XMEANT constraints in ITG alignment produces a more semantically correct alignment and thus yields interesting improvements compared to conventional ITG alignment and to the traditional GIZA++ alignment.

Finally, we also tested the performance of our model against MOSES hierarchical and MOSES phrase based translation baselines. We observed that systems using our semantically based approach for word alignment are comparable to BITG alignment systems in terms of edit distance metrics like TER, WER, PER and CDER, and that they both highly outperform the GIZA++ alignment based system results for Chinese to English translations.

7. REFERENCES

[1] Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu, "Fully automatic semantic MT evaluation," in *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

[2] Chi-kiu Lo, Karteek Addanki, Markus Saers, and Dekai Wu, "Improving machine translation by training against an automatic semantic frame based evaluation metric," in *51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, 2013.

[3] Chi-kiu Lo and Dekai Wu, "Can informal genres be better translated by tuning on automatic semantic metrics?," in *14th Machine Translation Summit (MT Summit XIV)*, 2013.

[4] Chi-kiu Lo, Meriem Beloucif, and Dekai Wu, "Improving machine translation into Chinese by tuning against Chinese MEANT," in *International Workshop on Spoken Language Translation (IWSLT 2013)*, 2013.

[5] Meriem Beloucif, Chi kiu Lo, and Dekai Wu, "Improving meant based semantically tuned smt," in *11 th International Workshop on spoken Language Translation (IWSLT 2014), 34-41 Lake Tahoe, California*, 2014.

[6] Chi-kiu Lo, Meriem Beloucif, Markus Saers, and Dekai Wu, "XMEANT: Better semantic MT evaluation without reference translations," in *52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, 2014.

[7] Karteek Addanki, Chi-kiu Lo, Markus Saers, and Dekai Wu, "LTG vs. ITG coverage of cross-lingual verb frame alternations," in *16th Annual Conference of the European Association for Machine Translation (EAMT-2012)*, Trento, Italy, May 2012.

[8] Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederik Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, "A statistical approach to machine translation," *Computational Linguistics*, vol. 16, no. 2, pp. 79–85, 1990.

[9] Stephan Vogel, Hermann Ney, and Christoph Tillmann, "HMM-based word alignment in statistical translation," in *The 16th International Conference on Computational linguistics (COLING-96)*, 1996, vol. 2, pp. 836–841.

[10] Dekai Wu, "Stochastic inversion transduction grammars and bilingual parsing of parallel corpora," *Computational Linguistics*, vol. 23, no. 3, pp. 377–403, 1997.

[11] Markus Saers and Dekai Wu, "Improving phrase-based translation via word alignments from stochastic inversion transduction grammars," in *Third Workshop on Syntax and Structure in Statistical Translation (SSST-3)*, Boulder, Colorado, June 2009, pp. 28–36.

[12] Markus Saers, Joakim Nivre, and Dekai Wu, "Word alignment with stochastic bracketing linear inversion transduction grammar," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, Los Angeles, California, June 2010, pp. 341–344.

[13] Markus Saers, Joakim Nivre, and Dekai Wu, "Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm," in *11th International Conference on Parsing Technologies (IWPT'09)*, Paris, France, October 2009, pp. 29–32.

[14] Richard Zens and Hermann Ney, “A comparative study on reordering constraints in statistical machine translation,” in *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Stroudsburg, Pennsylvania, 2003, pp. 144–151.

[15] Hao Zhang and Daniel Gildea, “Stochastic lexicalized inversion transduction grammar for alignment,” in *43rd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, Ann Arbor, Michigan, June 2005, pp. 475–482.

[16] Adrià De Gispert, Deepa Gupta, Maja Popovic, Patrik Lambert, Jose B.Marino, Marcello Federico, Hermann Ney, and Rafael Banchs, “Improving statistical word alignment with morpho-syntactic transformations,” in *Advances in Natural Language Processing*, 2006, pp. 368–379.

[17] Chris Dyer, Jonathan Clark, Alon Lavie, and Noah A.Smith, “Unsupervised word alignment with arbitrary features,” in *49th Annual Meeting of the Association for Computational Linguistics*, 2011.

[18] Jeff Ma, Spyros Matsoukas, and Richard Schwartz, “Improving low-resource statistical machine translation with a novel semantic word clustering algorithm,” in *Proceedings of the MT Summit XIII*, 2011.

[19] Theerawat Songyot and David Chiang, “Improving word alignment using word similarity,” in *52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

[20] Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky, “Shallow semantic parsing using support vector machines,” in *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, 2004.

[21] Chi-kiu Lo and Dekai Wu, “MEANT: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles,” in *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, 2011.

[22] Chi-kiu Lo and Dekai Wu, “Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics,” in *Sixth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-6)*, 2012.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “BLEU: a method for automatic evaluation of machine translation,” in *40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, Pennsylvania, July 2002, pp. 311–318.

[24] George Doddington, “Automatic evaluation of machine translation quality using n-gram co-occurrence statistics,” in *The second international conference on Human Language Technology Research (HLT ’02)*, San Diego, California, 2002.

[25] Satanjeev Banerjee and Alon Lavie, “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments,” in *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, June 2005.

[26] Gregor Leusch, Nicola Ueffing, and Hermann Ney, “CDer: Efficient MT evaluation using block movements,” in *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

[27] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney, “A evaluation tool for machine translation: Fast evaluation for MT research,” in *The Second International Conference on Language Resources and Evaluation (LREC 2000)*, 2000.

[28] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, “A study of translation edit rate with targeted human annotation,” in *7th Biennial Conference Association for Machine Translation in the Americas (AMTA 2006)*, Cambridge, Massachusetts, August 2006, pp. 223–231.

[29] Meriem Beloucif, Markus Saers, and Dekai Wu, “Improving semantic smt via soft semantic role label constraints on itg alignments,” in *Machine Translation Summit XV, Miami, Florida*, 2015.

[30] Zhaojun Wu, Yongsheng Yang, and Pascale Fung, “C-ASSERT: Chinese shallow semantic parser,” 2006.

[31] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst, “Moses: Open source toolkit for statistical machine translation,” in *Interactive Poster and Demonstration Sessions of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, Prague, Czech Republic, June 2007, pp. 177–180.

[32] Andreas Stolcke, “SRILM – an extensible language modeling toolkit,” in *7th International Conference*

on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002), Denver, Colorado, September 2002, pp. 901–904.

[33] Omar F. Zaidan, “Z-MERT: A fully configurable open source tool for minimum error rate training of machine translation systems,” *The Prague Bulletin of Mathematical Linguistics*, vol. 91, pp. 79–88, 2009.

[34] Franz Josef Och, “Minimum error rate training in statistical machine translation,” in *41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan, July 2003, pp. 160–167.