Improving word alignment for low resource languages using English monolingual SRL

Meriem Beloucif, Markus Saers and Dekai Wu
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{mbeloucif|masaers|dekai}@cs.ust.hk

Abstract

We introduce a new statistical machine translation approach specifically geared to learning translation from low resource languages, that exploits monolingual English semantic parsing to bias inversion transduction grammar (ITG) induction. We show that in contrast to conventional statistical machine translation (SMT) training methods, which rely heavily on phrase memorization, our approach focuses on learning bilingual correlations that help translating low resource languages, by using the output language semantic structure to further narrow down ITG constraints. This approach is motivated by previous research which has shown that injecting a semantic frame based objective function while training SMT models improves the translation quality. We show that including a monolingual semantic objective function during the learning of the translation model leads towards a semantically driven alignment which is more efficient than simply tuning loglinear mixture weights against a semantic frame based evaluation metric in the final stage of statistical machine translation training. We test our approach with three different language pairs and demonstrate that our model biases the learning towards more semantically correct alignments. Both GIZA++ and ITG based techniques fail to capture meaningful bilingual constituents, which is required when trying to learn translation models for low resource languages. In contrast, our proposed model not only improve translation by injecting a monolingual objective function to learn bilingual correlations during early training of the translation model, but also helps to learn more meaningful correlations with a relatively small data set, leading to a better alignment compared to either conventional ITG or traditional GIZA++ based approaches.

1 Introduction

In this paper we introduce a new approach for inversion transduction grammar (ITG) induction for low resource languages. Our induction algorithm uses the output language (English) semantic frames. Recent research showed that including a semantic frame based objective function at an early stage of training statistical machine translation (SMT) systems helps to learn more meaningful word alignments (Beloucif *et al.*, 2015) rather than relying on tuning against a semantic based objective function such as MEANT (Lo *et al.*, 2012), which improves the translation adequacy (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b; Beloucif *et al.*, 2014). We show that integrating a semantic based objective function much earlier in the training pipeline not only helps to learn more semantically correct alignments, but also helps us get rid of the heavy memorization used in conventional training methods, which is paramount for low resource languages where data sparseness makes memorization ineffective.

Our approach is also motivated by the fact that inversion transduction grammar alignments have previously been empirically shown to cover 100% of crosslingual semantic frame alternations, while ruling out the majority of incorrect alignments (Addanki *et al.*, 2012). We experiment on three different language pairs from the DARPA LORELEI study on efficient learning under low resource conditions: Chinese, Hausa, Uzbek, always translating into English.

We show that integrating a semantic frame based objective function much earlier in the training pipeline not only produces more semantically correct alignments but also helps to learn bilingual correlations without memorizing from a huge amount of parallel corpora. We believe that low resource conditions are

Proceedings of the Sixth Workshop on Hybrid Approaches to Translation,
pages 51–60, Osaka, Japan, December 11, 2016.

more interesting than high resource conditions because they are both scientifically and socioeconomically more interesting as they emphasize issues of efficient generalization as opposed to mere memorization from big data collections. We report results and examples showing that this way for inducing ITGs gives better translation quality compared to the conventional ITG (Saers and Wu, 2009) and GIZA++ (Och and Ney, 2000) alignments.

2 Related work

2.1 Alignment

Word alignment is considered to be an important step in training machine translation 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 (Brown *et al.*, 1990), which are known to produce weak word alignment since they allow unstructured movement of words. Then use heuristics to combine alignments of both directions to produce the final alignment. A hidden Markov model (HMM) based alignment was proposed (Vogel *et al.*, 1996), 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 any random word-to-word permutations. Such an alignment generally hurts the translation accuracy. The traditional GIZA++ (Och and Ney, 2000) toolkit implements both IBM and HMM models described above.

Saers and Wu (2009) proposed a better method of producing word alignment by training inversion transduction grammars (Wu, 1997). One problem encountered with such a model was the exhaustive biparsing that runs in $O(n^6)$. A more efficient version that runs in $O(n^3)$ was proposed later (Saers *et al.*, 2009).

Zens and Ney (2003) show that ITG constraints allow a higher flexibility in word ordering for longer sentences than the conventional IBM model. Furthermore, they demonstrate 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. Zhang and Gildea (2005) presented a version of ITG where rule probabilities are lexicalized throughout the synchronous parse tree for efficient training which helped to align sentences up to 15 words.

Some of the previous work on word alignment used morphological and syntactic features (De Gispert *et al.*, 2006). Some loglinear models have been proposed to incorporate those features (Dyer *et al.*, 2011). The problem with those approaches is that they require language specific knowledge and that they work better on more morphologically rich languages.

Few studies that approximately integrate semantic knowledge in computing word alignment are proposed by Ma *et al.* (2011) and Songyot and Chiang (2014). However, the former needs to have a prior word alignment learned on lexical words. The authors in the latter model proposed a semantic oriented word alignment. However, the problem is, they need to extract word similarity from the monolingual data for both languages, which is problematic in low resource conditions, then produce alignments using word similarities.

2.2 Inversion transduction grammars

Inversion transduction grammars, or ITGs, (Wu, 1997) are by definition a subset of syntax-directed transduction grammar (Lewis and Stearns, 1968; Aho and Ullman, 1972). A transduction is a set of bisentences that define the relation between an input language L_0 and an output language L_1. Accordingly, transduction grammars are able to:

$$\begin{cases} generate & (e, f \mid S) \\ translate & (e \mid f, S) \text{ or } (f \mid e, S) \\ accept & (S \mid e, f) \end{cases} \tag{1}$$

Table 1: The size of the different data sets in sentence pairs (foreign-English).

	Uzbek	Hausa	Chinese
Training	148,190	76,910	39,953
Development	1,200	1,000	1,512
Test	600	500	489

where $(e\ ,\ f)$is a sentence pair in L_0 and L_1 and S is the start symbol. Inversion transductions are syntax-directed transductions generated by inversion transduction grammars.

An ITG can always be written in a 2-normal form. Representing the ITG as a tuple $\langle N, V_0, V_1, R, S\rangle$ where N is a set of nonterminals, V_0 and V_1 are the tokens of L_0 and L_1 respectively, R is a set of transduction rules and $S \in N$ is the start symbol, each transduction rule can be restricted to one of the following forms:

$$\begin{gathered} S \rightarrow A \\ A \rightarrow [BC] \\ A \rightarrow \langle BC\rangle \\ A \rightarrow e/\epsilon \\ A \rightarrow \epsilon/f \\ A \rightarrow e/f \end{gathered}$$

where S, A, B, C are the non-terminals, e, f are tokens in the two languages and ϵ is the empty token.

ITGs allow both straight and inverted rules, straight transduction rules use square brackets and take the form $A \rightarrow [BC]$ and inverted rules use inverted brackets and take the form $A \rightarrow \langle BC\rangle$. Straight transduction rules generate transductions with the same order in L_0 and L_1 , inverted rules on the other hand, generate transduction in an inverted order. This means that, in the parse tree, the children instantiated by straight rules are read in the same order and children instantiated in an inverted order are read in an inverted order in L_1.

The rule probability function p is initialized using uniform probabilities for the structural rules, and a translation table t that is trained using IBM model 1 (Brown *et al.*, 1993) in both directions.

There are also many ways to formulate the model over ITGs: Wu (1995); Zhang and Gildea (2005); Chiang (2007); Cherry and Lin (2007); Blunsom *et al.* (2009); Haghighi *et al.* (2009); Saers *et al.* (2010); Neubig *et al.* (2011).

In this work, we use BITGs or bracketing transduction grammars (Saers *et al.*, 2009) which only use one single nonterminal category and surprisingly achieve good results.

2.3 Semantic frames in the MT training pipeline

Semantic role labeling (SRL) is an important task in natural language processing since it helps to define the basic event structure in a given sentence: *who did what to whom*, *for whom*, *when*, *where*, *how* and *why* as defined in (Pradhan *et al.*, 2004; Lo and Wu, 2011, 2012; Lo *et al.*, 2012). This approach gives a better way of understanding the meaning of a given sentence than the conventional syntax-based parsing.

Recent approaches in semantic role labeling use unsupervised machine learning techniques to automatically find the semantic roles. They generally use FrameNet (Gildea and Jurafsky, 2002) or Proposition Bank (Palmer *et al.*, 2005) notation to specify what a predicate is and what the other arguments are. The most recent research that include SRL in the SMT pipeline was done for MT evaluation. The MEANT family of metrics are semantic evaluation metrics that correlate more closely with human adequacy judgements than the commonly used surface based metrics (Lo and Wu, 2011, 2012; Lo *et al.*, 2012; Lo and Wu, 2013b; Macháček and Bojar, 2013).

Unlike *n*-gram or edit-distance based metrics, the MEANT family of metrics (Lo and Wu, 2011, 2012; Lo *et al.*, 2012) 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 (Pradhan *et al.*, 2004). Recent works have shown that the semantic frame based metric, MEANT, correlates better with human adequacy judgment than common evaluation metrics (Lo and Wu, 2011, 2012;

Algorithm Token based ITG-indcution and alignment.

C ▷ The parallel corpus
c ▷ The rule counts
$G = \langle N, W_0, W_1, R, S \rangle$ ▷ The empty ITG
$A \in N$ ▷ The bracketing symbol
p ▷ The rule probability function to estimate
a ▷ The alignments
$sum \leftarrow 0$ ▷ The sum of all counts
$R \leftarrow R \cup \{S \to A, A \to [AA], A \to \langle AA \rangle\}$
$p(S \to A) = 1$
$p(A \to [AA]) = \frac{1}{4}$
$p(A \to \langle AA \rangle) = \frac{1}{4}$
for parallel sentences $e_{0..T}/f_{0..V} \in C$ **do**
  for $0 \leq s < T$ **do**
    $W_0 \leftarrow W_0 \cup \{e_{s..s+1}\}$
    $R \leftarrow R \cup \{A \to e_{s..s+1}/\epsilon\}$
    $c_{A \to e_{s..s+1}/\epsilon} \leftarrow c_{A \to e_{s..s+1}/\epsilon} + 1$
    $sum \leftarrow sum + 1$
  for $0 \leq u < V$ **do**
    $W_1 \leftarrow W_1 \cup \{f_{u..u+1}\}$
    $R \leftarrow R \cup \{A \to \epsilon/f_{u..u+1}\}$
    $c_{A \to \epsilon/f_{u..u+1}} \leftarrow c_{A \to \epsilon/f_{u..u+1}} + 1$
    $sum \leftarrow sum + 1$
  for $0 \leq s < T$ **do**
    for $0 \leq u < V$ **do**
      $R \leftarrow R \cup \{A \to e_{s..s+1}/f_{u..u+1}\}$
      $c_{A \to e_{s..s+1}/f_{u..u+1}} \leftarrow c_{A \to e_{s..s+1}/f_{u..u+1}} + 1$
      $sum \leftarrow sum + 1$
for rule $A \to e/f \in R$ **do**
  $p(A \to e/f) \leftarrow \frac{1}{2} \frac{c_{A \to e/f}}{sum}$
repeat
  $p \leftarrow reestimate\_with\_em(G, p, C)$
until convergence
for parallel sentences $e_{0..T}/f_{0..V} \in C$ **do**
  $a_{e_{0..T}/f_{0..V}} \leftarrow viterbi\_parse(G, p, e_{0..T}/f_{0..V})$
return a

Figure 1: Token based BITG induction algorithm.

Table 2: Tuning the error penalty on the Chinese-English translation set.

	cased/uncased					
Weight	BLEU	METEOR	TER	WER	PER	CDER
0	16.29/16.63	36.9/38.9	69.09/68.69	71.34/71.03	60.78/60.22	67.89/67.44
0.01	15.93/16.34	36.4/38.6	69.14/68.77	71.80/71.42	60.99/60.43	68.29/67.87
0.1	15.77/15.99	37.0/38.9	69.30/68.90	71.85/71.48	60.46/59.90	68.18/67.76
0.5	16.90/17.19	37.9/40.1	68.85/68.53	71.53/71.26	60.14/59.61	67.44/67.18
0.6	**17.06/17.38**	**38.0/40.1**	**68.69/68.32**	**71.48/71.16**	**59.87/59.34**	**67.47/67.12**
0.9	16.34/16.60	37.4/39.3	69.80/69.33	72.33/71.96	60.75/60.19	68.58/68.18

Lo *et al.*, 2012) such as BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006). 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 (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b; Beloucif *et al.*, 2014) significantly improves the quality of the MT output. Beloucif *et al.* (2015) showed 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.

Table 3: Tuning the error penalty on the Hausa-English translation set.

	cased/uncased					
Weight	BLEU	METEOR	TER	WER	PER	CDER
0	16.60/17.14	44.8/47.8	70.63/69.69	73.16/72.46	58.24/56.77	69.59/68.71
0.01	16.83/17.37	43.9/46.7	71.06/70.08	73.62/72.85	58.96/57.36	70.05/69.02
0.1	17.35/17.87	44.6/47.6	69.99/69.05	72.65/71.93	58.17/56.59	69.10/68.08
0.5	17.10/17.57	44.2/47.2	70.39/69.50	72.92/72.19	58.92/57.47	69.45/68.49
0.6	**17.44/17.98**	**45.0/47.9**	**69.94/68.92**	**72.47/71.77**	**58.18/56.70**	**68.92/67.97**
0.9	16.99/17.49	44.9/48.0	70.18/69.21	72.78/56.55	58.08/56.55	69.17/68.24

3 Semantic frame based ITG induction for low resource languages

3.1 Word alignment

We implement a token based BITG system as our ITG baseline. Our choice of BITG constraints is based on previous work that has shown that BITG based alignments outperformed GIZA++ alignments (Saers *et al.*, 2009).

Figure 1 shows the BITG induction algorithm that we use in this paper. 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 contains 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 (Saers *et al.*, 2009) with a beam width of 100. In the last iteration, we extract the alignments imposed by the Viterbi parses as the word alignments outputted by the system.

Our proposed model injects a monolingual semantic frame based objective function into the BITG induction phase. We introduce an error weight between 0 and 1, that the inside probability is multiplied by if the English side of a bispan crosses any of the spans in the English SRL parse. The details of the approach are as follows:

$$\alpha' = \begin{cases} \alpha_{A_{s,t,u,v}} \times c_0 & \text{if } \forall_{(i,j)} \quad i \le s \wedge j \le s, \\ & \qquad\quad s \le i \wedge j \le t, \\ & \qquad\quad t \le i \wedge t \le j, \\ & \qquad\quad i \le s \wedge t \le j, \\ \alpha & \text{otherwise} \end{cases} \tag{2}$$

where α represents the inside probabliity, α' is the new estimated inside probability, (s, t) are the output language sentence spans, (i, j) are the English SRL parse spans. To ensure that we are not testing on any training data, we are doing something unusual: we tune the error weights on two different languages, and then test the best error weight on a third language. To test our method on Uzbek-English translations, we first tune the error weights using two language pairs: Chinese-English and Hausa-English translation. For both language pairs, we tune the error weights via grid search. Tables 2 and 3 represent the results that we got by experimenting with different error weights in both Chinese-English and Hausa-English test sets respectively. The best error weight that we got from both tunings equals to 0.6. We then apply the optimized selected weight to train an Uzbek-English translation model. This error weight is multiplied by the inside probabilities α during the BITG training if the English side of the ITG bispan crosses the English SRL parse as described in the function above.

We also train 10 iterations of EM of the new model and use Viterbi parsing to extract the alignments. We contrast the performance of our proposed monolingual semantic frame based alignment to the conventional BITG alignment and to the traditional GIZA++ baseline with grow-diag-final-and to harmonize both alignment directions.

Table 4: Translation quality of an Uzbek-English phrase based SMT system build on three different alignment methods.

	cased/uncased					
Alignments	BLEU	METEOR	TER	WER	PER	CDER
GIZA++	16.28/17.09	40.7/42.8	82.20/80.91	88.51/87.71	66.70/64.61	79.47/78.11
BITG	16.85/17.66	38.8/40.9	79.75/78.12	85.53/84.60	65.04/62.89	76.93/75.51
Monolingual English SRL	**17.40/18.15**	**41.0/43.4**	**79.25/77.72**	**85.20/84.48**	**63.29/61.13**	**76.36/75.00**

Input
Mamlakatimizga tashrif buyurgan Indoneziya Respublikasi tashqi ishlar vaziri Hasan Virayuda 13 may kuni O'zbekiston Respublikasi Oliy Majlisi Qonunchilik palatasi Spikeri Dilorom Toshmuhamedova bilan uchr ashdi

Ref
Foreign Minister of Indonesia Hasan Wirayuda met Speaker of the Legislative Chamber of Oliy Majlis of Uzbekistan Dilorom Tashmuhamedova on 13 May .

Giza++
is on a visit in Uzbekistan Minister of Foreign Affairs of the Republic of Indonesia Hasan Wirayuda said on 13 May , he met the Speaker of the Legislative Chamber of Oliy Majlis of Uzbekistan Dilorom Tashmuhamedova

BITG
Members of the delegation , headed by the Minister of Foreign Affairs of the Republic of Indonesia Hasan Wirayuda on May 13 , she met the Speaker of the Legislative Chamber of Oliy Majlis of Uzbekistan Dilorom Tashmuhamedova .

Proposed model
the Minister of Foreign Affairs of the Republic of Indonesia Hasan Wirayuda on 13 May , he met the Speaker of the Legislative Chamber of Oliy Majlis of Uzbekistan Dilorom Tashmuhamedova.

Figure 2: An example extracted from the test data for the Uzbek-English translations.

3.2 Baseline

Our experiments are part of the DARPA LORELEI study on efficient learning under low resource conditions therefore we purposely use relatively small corpora in different languages. We tried to show that including semantic frames earlier in learning SMT systems can help us to learn from relatively small corpora, in contrast to traditional SMT training models, which require expensive huge corpora. Table 1 represents the size of the three datasets used for our experimental setup. We tried to vary the data size and the language family for tuning the error weight and testing our proposed model to show that our approach is not language dependent and can easily be generalized across languages. We adopted the DARPA LORELEI program approach by using a relatively small Chinese corpus, a medium Hausa corpus and a slightly larger Uzbek corpus, we show that our approach is able to learn from small to medium datasets and does not rely on heavy memorization.

We tested the different alignments described above by using the standard MOSES toolkit (Koehn *et al.*, 2007), and a 4-gram language model learned with the SRI language model toolkit (Stolcke, 2002) trained on the training data of each language respectively. To tune the loglinear mixture weights, we use *k*-best MIRA (Cherry and Foster, 2012), a version of margin-based classification algorithm or MIRA (Chiang, 2012).

4 Results

We compared the performance of the semantic frame based BITG alignments against both the conventional token based BITG alignments and the traditional GIZA++ alignments. We evaluated our MT output using the surface based evaluation metrics BLEU (Papineni *et al.*, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006). Table 4 represents the result of testing our approach with the best tuned weight on Uzbek-English translations. We see that the alignment based on our proposed algorithm helps to achieve much higher scores across all metrics in comparison to both conventional BITG and GIZA++ alignments.

Figure 2 shows an interesting example extracted from the Uzbek-English translations, and compares the performance of our proposed model to both a GIZA++ based model and a BITG based model. We notice that our proposed model gives the output that best reflects the meaning of the sentence according to the reference translation. GIZA++ gives a relatively bad translation. BITG based model mixes the gender of "the prime minister Hasan Wirayuda" and refers to him by "she" instead of "he". Our proposed model on the other hand, is able to capture the general meaning of the sentence, and produces a relatively fluent output in comparison to both GIZA++ and BITG.

The results and examples we see above show that we should be more focused on incorporating semantic information during the actual early stage learning of the structure of the translation model, rather than merely tuning a handful of late stage loglinear mixture weights against a semantic objective function.

5 Conclusion

In this paper we have presented a semantically driven alignment method for low resource languages, where we use an English monolingual semantic frame parse and translation lexicons for BITG induction. We have shown that including a semantic frame based objective function at an early stage of learning SMT training helps to improve the quality of the MT translation for low resource languages. We experimented on three different language pairs from the DARPA LORELEI study on efficient learning under low resource conditions and have demonstrated that using a semantic frame based objective function during the actual learning of the translation model helps to learn good bilingual correlations with a relatively small dataset in contrast to conventional SMT systems.

Finally, we have shown that our proposed system produces a more semantically correct alignment and thus yields an improvement in comparison to the conventional BITG alignments and to the traditional GIZA++ alignments.

References

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.

Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation, and Compiling*. Prentice-Halll, Englewood Cliffs, New Jersey, 1972.

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.

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.

Meriem Beloucif, Markus Saers, and Dekai Wu. Improving semantic smt via soft semantic role label constraints on itg alignments. In *Machine Translation Summit XV (MT Summit 2015)*, pages 333–345, Miami, USA, October 2015.

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. A Gibbs sampler for phrasal synchronous grammar induction. In *Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, pages 782–790, Suntec, Singapore, August 2009.

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*, 16(2):79–85, 1990.

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

Colin Cherry and George Foster. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the ACL: Human Language Technologies*. Association for Computational Linguistics, 2012.

Colin Cherry and Dekang Lin. Inversion transduction grammar for joint phrasal translation modeling. In *Syntax and Structure in Statistical Translation (SSST)*, pages 17–24, Rochester, New York, April 2007.

David Chiang. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228, 2007.

David Chiang. Hope and fear for discriminative training of statistical translation models. *The Journal of Machine Learning Research*, 13:1159–1187, April 2012.

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*, pages 368–379, 2006.

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.

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.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. Better word alignments with supervised ITG models. In *Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP 2009)*, pages 923–931, Suntec, Singapore, August 2009.

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)*, pages 177–180, Prague, Czech Republic, June 2007.

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.

Philip M. Lewis and Richard E. Stearns. Syntax-directed transduction. *Journal of the Association for Computing Machinery*, 15(3):465–488, 1968.

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.

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.

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.

Chi-kiu Lo and Dekai Wu. MEANT at WMT 2013: A tunable, accurate yet inexpensive semantic frame based mt evaluation metric. In *8th Workshop on Statistical Machine Translation (WMT 2013)*, 2013.

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

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.

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.

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.

Matouš Macháček and Ondřej Bojar. Results of the WMT13 metrics shared task. In *Eighth Workshop on Statistical Machine Translation (WMT 2013)*, Sofia, Bulgaria, August 2013.

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. An unsupervised model for joint phrase alignment and extraction. In *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 632–641, Portland, Oregon, June 2011.

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.

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *The 38th Annual Meeting of the Association for Computational Linguistics (ACL 2000)*, pages 440–447, Hong Kong, October 2000.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

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)*, pages 311–318, Philadelphia, Pennsylvania, July 2002.

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.

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)*, pages 28–36, Boulder, Colorado, June 2009.

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)*, pages 29–32, Paris, France, October 2009.

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)*, pages 341–344, Los Angeles, California, June 2010.

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)*, pages 223–231, Cambridge, Massachusetts, August 2006.

Theerawat Songyot and David Chiang. Improving word alignment using word similarity. In *52nd Annual Meeting of the Association for Computational Linguistics*, 2014.

Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)*, pages 901–904, Denver, Colorado, September 2002.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *The 16th International Conference on Computational linguistics (COLING-96)*, volume 2, pages 836–841, 1996.

Dekai Wu. Trainable coarse bilingual grammars for parallel text bracketing. In *Third Annual Workshop on Very Large Corpora (WVLC-3)*, pages 69–81, Cambridge, Massachusetts, June 1995.

Dekai Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377–403, 1997.

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)*, pages 144–151, Stroudsburg, Pennsylvania, 2003.

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