Improving Semantic SMT via Soft Semantic Role Label Constraints on ITG Alignments

Meriem BELOUCIF mbeloucif@cs.ust.hk
Markus SAERS masaers@cs.ust.hk
Dekai WU dekai@cs.ust.hk
Human Language Technology Center
Hong Kong University of Science and Technology, Hong Kong

Abstract

We show that applying semantic role label constraints to bracketing ITG alignment to train MT systems improves the quality of MT output in comparison to the conventional BITG and GIZA alignments. Moreover, we show that applying soft constraints to SRL-constrained BITG alignment leads to a better translation system compared to using hard constraints which appear too harsh to produce meaningful biparses. We leverage previous work demonstrating that BITG alignments are able to fully cover cross-lingual semantic frame alternations, by using semantic role labeling to further narrow BITG constraints, in a soft fashion that avoids losing relevant portions of the search space. SRL-based evaluation metrics like MEANT have shown that tuning towards preserving the shallow semantic structure across translations, robustly improves translation performance. Our approach brings the same intuition into the training phase. We show that our new alignment outperforms both conventional Moses and BITG alignment baselines in terms of the adequacy-oriented MEANT scores, while still producing comparable results in terms of edit distance metrics.

1 Introduction

The quality of machine translation output relies heavily on word alignment. However, the most widespread approach to word alignment is the *ad hoc* method of training IBM models (Brown *et al.*, 1990) in both directions and combining their results using various heuristics. Word alignments based on inversion transduction grammars or ITGs (Wu, 1997), on the other hand, provide a more structured model leading to efficient and optimal bidirectional alignments.

In this paper we introduce an improved word aligner based on applying soft semantic role label constraints to ITG alignment. We show that both translation adequacy and fluency can be improved by replacing the conventional GIZA++ based alignment (Och and Ney, 2000) with more semantically motivated alignments obtained through training

ITGs (Saers and Wu, 2009) under soft SRL constraints. The new approach is motivated by Addanki *et al.* (2012) who demonstrated empirically that the semantic role reorderings found in cross-lingual SRL frames is essentially 100% covered by ITG constraints, which suggests that it should be possible to use ITG constraints as a starting point under which to align semantic frames as we do in this paper.

Our approach is further motivated by the fact that including semantic role labeling in the SMT pipeline in a different way has already been shown to increase translation quality. The semantic frame based evaluation metric MEANT, which was shown to correlate better with human adequacy judgment than conventional surface based evaluation metrics (Lo *et al.*, 2012), can be used as an objective function for tuning SMT. Tuning to MEANT, which attempts to optimize the degree to which a sentence's semantic frames can be preserved across translation, was shown to improve translation quality across many metrics (Lo *et al.*, 2013b; Beloucif *et al.*, 2014). We show in this paper that including soft constraints based on semantic role labeling into the alignment training step yields both higher adequacy-oriented MEANT and, while still producing comparable scores on surface based and edit distance metrics.

2 Related work

2.1 Alignment

For most recent automatic machine translation systems, learning a good word alignment is paramount for producing meaningful translation. Unfortunately, conventional alignment algorithms such as IBM models (Brown *et al.*, 1990) and the HMM-alignment model (Vogel *et al.*, 1996) are flat and directed, meaning that (a) they allow unstructured movement of words leading to weak word alignment, (b) translations in one direction are considered in isolation, and (c) two separate alignments are needed to form a single *bidirectional alignment*. The harmonization of two directed alignments is typically done heuristically, which means that there is no model that considers the final bidirectional alignment that the translation system is trained on to be optimal. Transduction grammars, on the other hand, do provide a model that (a) is inherently *structurally compositional*, and (b) can provide *optimal bidirectional alignments*. Although this structured optimality comes at a higher cost in terms of time complexity, it allows for preexisting structured information to be incorporated into the model, and for models to be compared in a meaningful way.

There are different classes of transduction grammars, ranging from finite-state transduction grammar, via linear transduction grammar (Saers *et al.*, 2010) and inversion transduction grammar (Wu, 1997; Saers and Wu, 2009; Saers *et al.*, 2009), to syntax-directed transduction grammar (Lewis and Stearns, 1968; Aho and Ullman, 1972) and many ways to formulate the model over them: 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 paper, we introduce a semantically biased version of inversion transduction grammars (Wu, 1997) that is biased towards constituents that conform to monolingual semantic parses on the input and/or output languages, and compare their performance against (a) ITGs without such a bias, and (b) the conventional heuristics.

2.2 Semantic role labeling in MT

Our alignment method is fully compatible with the principle that a good translation is one where a human can successfully understand the main meaning of the output sentence as captured by the basic event structure: "*who did what to whom, when, where and why*" (Pradhan *et al.*, 2004; Lo and Wu, 2011, 2012; Lo *et al.*, 2012). The MEANT family of metrics are semantic evaluation MT evaluation metrics that correlate with human adequacy judgements more closely than most commonly used surface based metrics (Lo and Wu, 2011, 2012; Lo *et al.*, 2012; Lo and Wu, 2013b; Macháček and Bojar, 2013). MEANT compares the MT output sentence against provided reference translations, and produce a score measuring 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 (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b; Beloucif *et al.*, 2014). We show here, that soft incorporation of SRL constraints much earlier in the pipeline, at the word alignment stage of SMT training, can further improve translation adequacy.

2.3 Inversion transduction grammars

A transduction represents a set of bi-sentences that define the relation between an input language L_0 and an output language L_1. Accordingly, a transduction grammar generates a transduction or a set of bi-sentences, translates between sentences in L_0 and sentences in L_1, and accepts the sentence pairs in the transduction. Inversion transductions are a subset of syntax-directed transductions which are generated and parsed by inversion transduction grammars or ITGs (Wu, 1997). An ITG can always be written in 2-normal form and is represented by a tuple $\langle N, V_0, V_1, R, S\rangle$ where N is a set of non-terminals, V_0 and V_1 are the vocabularies of L_0 and L_1 respectively, R is a set of transduction rules and $S \in N$ is the start symbol. In the 2-normal form, each inversion transduction must be on one of the following forms:

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

Input	这 个 在 日本 还 没有 出售 。
Gloss	this one in Japan yet haven’t sell .
Reference	They do not sell this in Japan , yet .
GIZA++	this in Japan yet on sale .
ITG	this in Japan has not on sale .
ITG with hard SRL constraints	
ITG with soft SRL constraints	this is not sold in Japan yet .

Figure 1: MT output of the three different systems for a given Chinese sentence

	请	告诉	我去	旧金山	最	快	的	方法	。
please	■								
tell		■							
me		■							
the					■		■		
fastest					■	■			
way								■	
to			■						
go			■						
to			■						
San			■	■					
Francisco				■					
.									■

Figure 2: An alignment of a bisentence produced by GIZA++ alignment, with grow-diag-final-and as a heuristic

$$A \to e/f$$

ITGs allow straight and inverted rules such that straight transduction rules use square brackets and take the form $A \to [BC]$ and inverted rules use inverted brackets and take the form $A \to \langle BC \rangle$. Straight transduction rules generate transductions with the same order in L_0 and L_1 which means that, in the parse tree, the children instantiated by straight rules are read in the same order. Inverted transduction rules on the other hand, generate transductions with inverted order in L_0 and L_1, so the children instantiated by inverted rules are read left-to-right in L_0 and right-to-left in L_1. In this paper we show that an ITG-based SMT system is able to perform better on semantic metrics when biased towards respecting semantic role labeling.

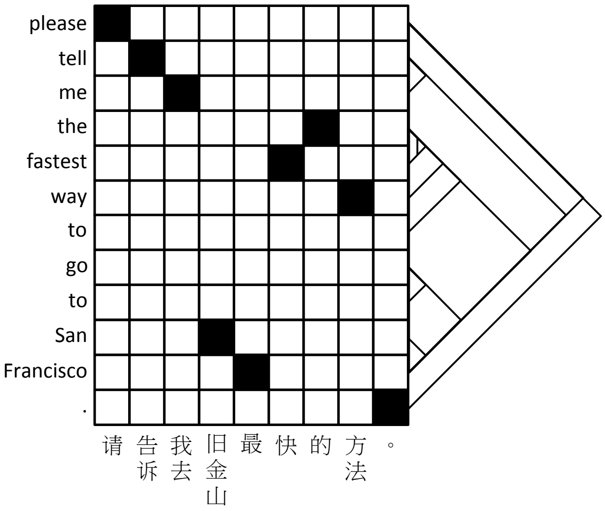

Figure 3: An alignment of a bisentence produced by ITG alignment

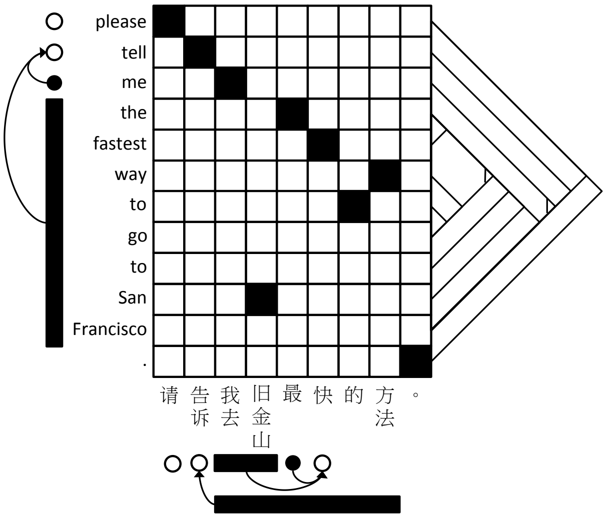

Figure 4: An alignment of a bisentence produced by ITG alignment using automatic shallow semantic parsing constraints. The input is parsed by a Chinese automatic shallow semantic parser.

3 SRL-constrained ITGs

The model we propose in this paper introduces soft constraints to bias a bracketing inversion transduction grammar (BITG) towards preferring bilingual constituents that conform to automatically generated monolingual semantic role labels on both sides. Because of the structural differences between monolingual SRLs and the bilingual BITG parses, we implement this as a penalty for BITG constituents violating the semantic role labels in either language.

The semantic roles and their fillers in a sentence sometimes span multiple syntactic units, or in technical terms: the semantic trees are (a) not necessarily consistent with the syntactic trees, and (b) not necessarily projective. Since BITG trees are defined to be projective, applying even a single monolingual SRL parse as a hard constraint would rule out *all* possible BITG trees, and *all* possible alignments for that sentence pair, since no BITG parse can conform to a non projective constraint. Even when the monolingual SRL trees are projective in both languages, there is a risk of overly constraining the search, as the only way for the BITG parser to satisfy incompatible SRL constraints is to sacrifice the lexical correspondences; the only way to conform to (a) the input language SRL, (b) the output language SRL, and (c) the ITG constraint may be to delete a constituent in the input language and insert it in the output language, even when there is a good translation between them, because translating them would violate at least one of the two constraints. The ITG constraint is what allows us to do this processing in polynomial time, so it is non-negotiable. As the lexical relation is what defines the word alignment, which is what we are interested in, we opt to soften the SRL constraints. In practice, the automatic SRL parses are fairly noisy, an engineering reason to soften them, but even with perfect SRL parsers, soft constraints are theoretically necessary.

The soft constraints takes the form of a fixed penalty that is paid whenever the BITG parser wants to introduce a bi-constituent that crosses a semantic constituent (the string a predicate or one of its role fillers span). No penalty is paid for bi-constituents completely covering a semantic constituent or by bi-constituents that are completely covered by semantic constituents. To allow for some degrees of freedom, we allows for two separate penalties, one for crossing an input language semantic constituent, λ_1, and one for crossing an output language semantic constituent, λ_0. These hyper parameters need to be set manually.

4 Experiment Setup

4.1 Data

For this paper, we tested our systems on Chinese to English translations. We used IWSLT 2007 data set for this experiment. The training set contains 39,953 sentences. The dev and test set contain 1512 sentences and test 489 sentences, respectively. Both the English and Chinese corpora were normalized for puntuation, tokenized and true-cased. We also used our own Chinese named entity recognition, and dedicated proper name translator.

4.2 Word alignment

We compare the performance of our SRL soft-constrained model to the SRL hard-constrained system and to the conventional unconstrained ITG baseline. We perform a grid search over soft-constraints hyper parameters to find the optimal settings. We then compared the performance of our proposed alignments to the conventional GIZA++ baseline with grow-diag-final-and to harmonize the two alignment directions.

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 (Saers *et al.*, 2009) with a beam width of 100. On the last iteration, we extract the alignments imposed by the Viterbi parses as the word alignments outputted by the system.

The new SRL-constrained ITG approach adds the crossing penalty based on automatic SRL parses discussed in Section 3 to the ITG baseline discussed above. The shallow semantic parses of the training data were produced using ASSERT (Pradhan *et al.*, 2004) and C-ASSERT (Wu *et al.*, 2006) for English and Chinese respectively.

We show how applying soft SRL constrained ITG alignment outperforms alignment both (a) without constraints and (b) with hard SRL constraints. The hyper parameter$\lambda_i = 0$ represents the hard constraints, $\lambda_i = 1$ represents the case with no constraints and λ_i between 0 and 1 are the soft constraints. We run some prior experiments and observed that applying hard SRL constraints did not lead to any alignment at all: the constraints were too harsh and did not permit any biparses. Soft SRL constrained ITGs, on the other hand, outperformed the unbiased BITG model in term of both adequacy-oriented MEANT scores. We noticed that $\lambda_0 = 1$ and $\lambda_1 = 0.5$ correspond to the best combination. The SRL constraints were only used during training of the probabilities of the ITG, and not when extracting the Viterbi parses and the corresponding word alignments.

4.3 SMT pipeline

To test the different alignments described in this paper, we used the standard Moses toolkit (Koehn *et al.*, 2007), with a 6-gram language model learned with the SRI language model toolkit (Stolcke, 2002), to train our baselines. We tested our approach using Moses hierarchical models. For tuning, we used ZMERT (Zaidan, 2009), a standard implementation of minimum error rate training or MERT (Och, 2003), we run each tuning task ten times for each system, then we decoded with both DEV and TEST set, we then chose the results according to what performed the best on the know development data. We also present the highest score achieved by each system among all runs (an upper bound on the score that can be achieved). We compared an edit-distance based metrics, BLEU (Papineni *et al.*, 2002) and TER (Snover *et al.*, 2006), and a semantic evaluation metric MEANT (Lo *et al.*, 2012) as the tuning objective.

5 Results

We compared the the performance of our soft SRL-constrained ITG alignment to (a) the GIZA++ baseline and (b) the unbiased BITG, for both BLEU, TER and MEANT tuned systems. We evaluated our MT output using the semantic metric MEANT (Lo *et al.*, 2012). Tables 1, 2 and 3 show the improvment in terms of MEANT scores for the SRL ITG aligned system in comparison to conventional ITG alignment and GIZA++ alignment for BLEU, TER and MEANT tuned systems respectively. The MEANT score for ITG based systems is considerably higher than the MEANT score for GIZA++ aligned model. We also observe that MEANT score for ITG with SRL constraints is better than the conventional ITG model. We believe that a better SRL-parser would yield a better system still. Tables 4, 5 and 6 give an upper bound on the results for the ten runs, we also see here that new semantically biased ITG model outperforms the baseline and ITG based alignment.

In addition, we evaluated the performance of our model using surface-based metrics such as 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), and observed that both the unbiased ITG and semantically biased ITG based systems yield comparable results that are high in comparison to conventional GIZA++ alignment.

Figure 1 shows an interesting example extracted from the test data. The Chinese input sentence has been pre-segmented into eight word-like units, and a word-for-word gloss reads approximately "this one in Japan yet haven't sell ." The translator who produce the reference translation took some liberties and introduced the actor *they*, who no doubt was present in the context, but is not needed when the sentence is considered in isolation. The conventional GIZA++ system fails completely, not only is the translation bad English, but the meaning it conveys is the polar opposite of the original sentence due to a dropped negation. Although these kind of errors are disastrous for the consumer of the translations, they are common with surface based systems, and not likely to be addressed with surface based tuning objectives. The unbiased ITG system produces a sentence that is understandable but far from good, comparable to something a second-language learner would write early on. The produced sentence does, however, convey the correct meaning, although some nuance is missing due to the dropped *yet*. The ITG system with *hard* SRL constraints outputs nothing, because the constraints prevented it from learning from most of the training examples. The ITG system with *soft* SRL constraints produces a sentence that conveys the correct meaning in good English. Considered in isolation, the produced sentence is even better than the man-made

Table 1: The optimal results given the known development set for the BLEU tuned systems

	cased						
System	MEANT	BLEU	METEOR	TER	WER	PER	CDER
Giza alignment	49.42	**25.07**	0.451	59.95	60.96	54.91	59.32
ITG alignment	50.02	24.13	**0.460**	**8.49**	**59.55**	53.98	**58.01**
SRL ITG alignment	**50.92**	23.67	**0.460**	58.78	59.79	**53.61**	58.25

Table 2: The optimal results given the known development set for the TER tuned systems

	cased						
System	MEANT	BLEU	METEOR	TER	WER	PER	CDER
Giza alignment	48.70	**23.02**	0.407	59.93	60.83	55.76	59.55
ITG alignment	49.76	21.72	0.431	**57.83**	**58.76**	53.90	**57.38**
SRL ITG alignment	**50.02**	21.92	**0.433**	58.78	59.55	**53.34**	58.01

Table 3: The optimal results given the known development set for the MEANT tuned systems

	cased						
System	MEANT	BLEU	METEOR	TER	WER	PER	CDER
Giza alignment	48.88	21.18	0.455	61.49	62.90	55.76	60.91
ITG alignment	49.12	**23.47**	**0.465**	60.30	61.20	55.36	59.53
SRL ITG alignment	**50.66**	22.83	0.446	**59.48**	**60.32**	**54.30**	**58.84**

reference translation, as it is closer to the original.

Figures 2-4 shows a sentence pair from the training data, and how the different systems end up aligning the words. The pre-segmented Chinese words correspond to *please*, *tell*, *I go*, *San Francisco*, *most*, *fast*, *of*, *way*, and full stop. The harmonized GIZA++ alignments (Figure 2) are good, except that *to go to San Francisco* and *the fastest* are grouped into atomic units, which means that this is not an example of *go*, *San Francisco*, or *fastest* being translated. The unconstrained ITG alignment (Figure 3) makes two mistakes: it insists on aligning 我去'I go' with *me* instead of *go*, and it aligns the superlative marker 最 with *Francisco*. The SRL constraints (Figure 4) are able to fix the latter, but not the former.

6 Conclusion

In this paper we showed that incorporating SRL constraints into the training of bracketing ITGs for early stage word alignment in SMT training leads to improved semantic

Table 4: The upper bound score among the eleven runs for BLEU tuned systems for each pipeline.

	cased						
System	MEANT	BLEU	METEOR	TER	WER	PER	CDER
Giza alignment	50.91	**25.07**	0.453	59.53	60.42	54.43	59.02
ITG alignment	50.75	24.41	0.469	**58.47**	**59.55**	53.90	**57.80**
SRL ITG alignment	**51.28**	24.67	**0.556**	58.78	59.69	**53.61**	58.25

Table 5: The upper bound score among the eleven runs for TER tuned systems for each pipeline.

	cased						
System	MEANT	BLEU	METEOR	TER	WER	PER	CDER
Giza alignment	49.94	23.02	0.408	59.40	60.52	55.58	59.14
ITG alignment	50.75	22.11	0.434	**57.16**	**58.57**	53.55	**57.30**
SRL ITG alignment	**50.92**	**24.70**	**0.441**	57.93	58.86	**53.34**	57.43

Table 6: The upper bound score among the eleven runs for MEANT tuned systems for each pipeline.

	cased						
System	MEANT	BLEU	METEOR	TER	WER	PER	CDER
Giza alignment	49.15	21.25	0.455	61.49	62.90	55.76	60.75
ITG alignment	49.94	**23.94**	**0.466**	60.30	61.20	55.36	59.53
SRL ITG alignment	**50.66**	22.83	0.457	**59.48**	**60.32**	**54.30**	**58.81**

translation adequacy. Moreover, we showed that applying soft SRL constraints in one of the languages produces better performance on the semantically oriented MEANT metric, in comparison to not applying any constraints. As automatically produced semantic parses for both languages are incompatible much of the time, we showed that the increased flexibility of soft constraints helps improve the word alignment quality. Finally, we observed that applying SRL constraints to BITG alignment using soft constraints not only improves MEANT scores but also retains the performance using surface oriented metrics like metrics like BLEU, METEOR, TER, WER, PER and CDER.

7 Acknowledgment

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract no. HR0011-12-C-0016, and GALE contract nos. HR0011-06-C-0022 and HR0011-06-C-0023; by the European Union under the FP7 grant agreement no. 287658; and by the Hong Kong Research Grants Council (RGC) research grants GRF620811, GRF621008, and GRF612806. 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. We are grateful to Pascale Fung, Yongsheng Yang and Zhaojun Wu for sharing the maximum entropy Chinese segmenter and C-ASSERT, the Chinese semantic parser, with us.

References

Karteek Addanki, Chi-kiu Lo, Markus Saers, and Dekai Wu. LTG vs. ITG coverage of cross-lingual verb frame alternations. In *EAMT-2012*, Trento, Italy, May 2012.

Alfred V. Aho and Jeffrey D. Ullman. *The Theory of Parsing, Translation, and Compiling*. Prentice-Halll, Englewood Cliffs, NJ, 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, MI, Jun 2005.

Meriem Beloucif, Chi kiu Lo, and Dekai Wu. Improving meant based semantically tuned smt. In *IWSLT 2014*, 2014.

Phil Blunsom, Trevor Cohn, Chris Dyer, and Miles Osborne. A Gibbs sampler for phrasal synchronous grammar induction. In *ACL-IJCNLP 2009*, pages 782–790, Suntec, Singapore, Aug 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.

Colin Cherry and Dekang Lin. Inversion transduction grammar for joint phrasal translation modeling. In *SSST*, pages 17–24, Rochester, NY, Apr 2007.

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

Aria Haghighi, John Blitzer, John DeNero, and Dan Klein. Better word alignments with supervised ITG models. In *ACL-IJCNLP 2009*, pages 923–931, Suntec, Singapore, Aug 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 *ACL 2007*, pages 177–180, Prague, Czech Republic, Jun 2007.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. CDer: Efficient MT evaluation using block movements. In *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 *ACL HLT 2011*, 2011.

Chi-kiu Lo and Dekai Wu. Unsupervised vs. supervised weight estimation for semantic MT evaluation metrics. In *SSST-6*, 2012.

Chi-kiu Lo and Dekai Wu. Can informal genres be better translated by tuning on automatic semantic metrics? In *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 *WMT 2013*, 2013.

Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. Fully automatic semantic MT evaluation. In *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 *ACL 2013*, 2013.

Chi-kiu Lo, Meriem Beloucif, and Dekai Wu. Improving machine translation into Chinese by tuning against Chinese MEANT. In *IWSLT 2013*, 2013.

Matouš Macháček and Ondřej Bojar. Results of the WMT13 metrics shared task. In *WMT 2013*, Sofia, Bulgaria, Aug 2013.

Graham Neubig, Taro Watanabe, Eiichiro Sumita, Shinsuke Mori, and Tatsuya Kawahara. An unsupervised model for joint phrase alignment and extraction. In *ACL HLT 2011*, pages 632–641, Portland, OR, Jun 2011.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. A evaluation tool for machine translation: Fast evaluation for MT research. In *LREC 2000*, 2000.

Franz Josef Och and Hermann Ney. Improved statistical alignment models. In *ACL 2000*, pages 440–447, Hong Kong, Oct 2000.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *ACL-2003*, pages 160–167, Sapporo, Japan, Jul 2003.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL-02*, pages 311–318, Philadelphia, PA, Jul 2002.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow semantic parsing using support vector machines. In *HLT-NAACL 2004*, 2004.

Markus Saers and Dekai Wu. Improving phrase-based translation via word alignments from stochastic inversion transduction grammars. In *SSST-3*, pages 28–36, Boulder, CO, Jun 2009.

Markus Saers, Joakim Nivre, and Dekai Wu. Learning stochastic bracketing inversion transduction grammars with a cubic time biparsing algorithm. In *IWPT'09*, pages 29–32, Paris, France, Oct 2009.

Markus Saers, Joakim Nivre, and Dekai Wu. Word alignment with stochastic bracketing linear inversion transduction grammar. In *NAACL HLT 2010*, pages 341–344, Los Angeles, CA, Jun 2010.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *AMTA 2006*, pages 223–231, Cambridge, MA, Aug 2006.

Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *ICSLP2002 - INTERSPEECH 2002*, pages 901–904, Denver, CO, Sep 2002.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. HMM-based word alignment in statistical translation. In *COLING-96*, volume 2, pages 836–841, 1996.

Zhaojun Wu, Yongsheng Yang, and Pascale Fung. C-ASSERT: Chinese shallow semantic parser, 2006.

Dekai Wu. Trainable coarse bilingual grammars for parallel text bracketing. In *WVLC-3*, pages 69–81, Cambridge, MA, Jun 1995.

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

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*, 91:79–88, 2009.

Hao Zhang and Daniel Gildea. Stochastic lexicalized inversion transduction grammar for alignment. In *ACL-05*, pages 475–482, Ann Arbor, MI, Jun 2005.