Better Semantic Frame Based MT Evaluation via Inversion Transduction Grammars

Dekai Wu **Lo Chi-kiu** **Meriem Beloucif** **Markus Saers**
HKUST
Human Language Technology Center
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
{jackielo|mbeloucif|masaers|dekai}@cs.ust.hk

Abstract

We introduce an inversion transduction grammar based restructuring of the MEANT automatic semantic frame based MT evaluation metric, which, by leveraging ITG language biases, is able to further improve upon MEANT's already-high correlation with human adequacy judgments. The new metric, called IMEANT, uses bracketing ITGs to biparse the reference and machine translations, but subject to obeying the semantic frames in both. Resulting improvements support the presumption that ITGs, which constrain the allowable permutations between compositional segments across the reference and MT output, score the phrasal similarity of the semantic role fillers more accurately than the simple word alignment heuristics (bag-of-word alignment or maximum alignment) used in previous version of MEANT. The approach successfully integrates (1) the previously demonstrated extremely high coverage of cross-lingual semantic frame alternations by ITGs, with (2) the high accuracy of evaluating MT via weighted f-scores on the degree of semantic frame preservation.

1 Introduction

There has been to date relatively little use of inversion transduction grammars (Wu, 1997) to improve the accuracy of MT evaluation metrics, despite long empirical evidence the vast majority of translation patterns between human languages can be accommodated within ITG constraints (and the observation that most current state-of-the-art SMT systems employ ITG decoders). We show that ITGs can be used to redesign the MEANT semantic frame based MT evaluation metric (Lo *et al.*, 2012) to produce improvements in accuracy and reliability. This work is driven by the motivation that especially when considering *semantic* MT metrics, ITGs would be seem to be a natural basis for several reasons.

To begin with, it is quite natural to think of sentences as having been generated from an abstract concept using a rewriting system: a stochastic grammar predicts how frequently any particular realization of the abstract concept will be generated. The bilingual analogy is a *transduction grammar* generating *a pair* of possible realizations of *the same* underlying concept. Stochastic transduction grammars predict how frequently a particular pair of realizations will be generated, and thus represent a good way to evaluate how well a pair of sentences correspond to each other.

The particular class of transduction grammars known as ITGs tackle the problem that the (bi)parsing complexity for general **syntax-directed transductions** (Aho and Ullman, 1972) is exponential. By constraining a syntax-directed transduction grammar to allow only monotonic **straight** and **inverted** reorderings, or equivalently permitting only binary or ternary rank rules, it is possible to isolate the low end of that hierarchy into a single equivalence class of **inversion transductions**. ITGs are guaranteed to have a two-normal form similar to context-free grammars, and can be biparsed in polynomial time and space ($O\left(n^6\right)$ time and $O\left(n^4\right)$ space). It is also possible to do approximate biparsing in $O\left(n^3\right)$ time (Saers *et al.*, 2009). These polynomial complexities makes it feasible to estimate the parameters of an ITG using standard machine learning techniques such as expectation maximization (Wu, 1995b) .

At the same time, inversion transductions have also been directly shown to be more than sufficient to account for the reordering that occur within semantic frame alternations (Addanki *et al.*, 2012). This makes ITGs an appealing alternative for eval-

Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 22–33,
October 25, 2014, Doha, Qatar.

uating the possible links between both semantic role fillers in different languages as well as the predicates, and how these parts fit together to form entire semantic frames. We believe that ITGs are not only capable of generating the desired structural correspondences between the semantic structures of two languages, but also provide meaningful constraints to prevent alignments from wandering off in the wrong direction.

In this paper we show that IMEANT, a new metric drawing from the strengths of both MEANT and inversion transduction grammars, is able to exploit bracketing ITGs (also known as BITGs or BTGs) which are ITGs containing only a single non-differentiated non terminal category (Wu, 1995a), so as to produce even higher correlation with human adequacy judgments than any automatic MEANT variants, or other common automatic metrics. We argue that the constraints provided by BITGs over the semantic frames and arguments of the reference and MT output sentences are essential for accurate evaluation of the phrasal similarity of the semantic role fillers.

In common with the various MEANT semantic MT evaluation metrics (Lo and Wu, 2011a, 2012; Lo *et al.*, 2012; Lo and Wu, 2013b), our proposed IMEANT metric measures the degree to which the basic semantic event structure is preserved by translation—the "who did what to whom, for whom, when, where, how and why" (Pradhan *et al.*, 2004)—emphasizing that a good translation is one that can successfully be understood by a human. In the other versions of MEANT, similarity between the MT output and the reference translations is computed as a modified weighted f-score over the semantic predicates and role fillers. Across a variety of language pairs and genres, it has been shown that MEANT correlates better with human adequacy judgment than both n-gram based MT evaluation metrics such as BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), and METEOR (Banerjee and Lavie, 2005), as well as edit-distance based metrics such as CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006) when evaluating MT output (Lo and Wu, 2011a, 2012; Lo *et al.*, 2012; Lo and Wu, 2013b; Macháček and Bojar, 2013). Furthermore, tuning the parameters of MT systems with MEANT instead of BLEU or TER robustly improves translation adequacy across different genres and different languages (English and Chinese) (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b). This has motivated our choice of MEANT as the basis on which to experiment with deploying ITGs into semantic MT evaluation.

2 Related Work

2.1 ITGs and MT evaluation

Relatively little investigation into the potential benefits of ITGs is found in previous MT evaluation work. One exception is **invWER**, proposed by Leusch *et al.* (2003) and Leusch and Ney (2008). The invWER metric interprets weighted BITGs as a generalization of the Levenshtein edit distance, in which entire segments (blocks) can be inverted, as long as this is done strictly compositionally so as not to violate legal ITG biparse tree structures. The input and output languages are considered to be those of the reference and machine translations, and thus are over the same vocabulary (say,English). At the sentence level, correlation of invWER with human adequacy judgments was found to be among the best.

Our current approach differs in several key respects from invWER. First,invWER operates purely at the surface level of exact token match, IMEANT mediates between segments of reference translation and MT output using lexical BITG probabilities.

Secondly, there is no explicit semantic modeling in invWER. Providing they meet the BITG constraints, the biparse trees in invWER are completely unconstrained. In contrast, IMEANT employs the same explicit, strong semantic frame modeling as MEANT, on both the reference and machine translations. In IMEANT, the semantic frames always take precedence over pure BITG biases. Compared to invWER, this strongly constrains the space of biparses that IMEANT permits to be considered.

2.2 MT evaluation metrics

Like invWER, other common surface-form oriented metrics like BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005; Denkowski and Lavie, 2014), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006) do not correctly reflect the meaning similarities of the input sentence. There are in fact several large scale meta-evaluations (Callison-Burch *et al.*, 2006; Koehn and Monz, 2006) reporting cases

where BLEU strongly disagrees with human judgments of translation adequacy.

Such observations have generated a recent surge of work on developing MT evaluation metrics that would outperform BLEU in correlation with human adequacy judgment (HAJ). Like MEANT, the TINE automatic recall-oriented evaluation metric (Rios *et al.*, 2011) aims to preserve basic event structure. However, its correlation with human adequacy judgment is comparable to that of BLEU and not as high as that of METEOR. Owczarzak *et al.* (2007a,b) improved correlation with human *fluency* judgments by using LFG to extend the approach of evaluating syntactic dependency structure similarity proposed by Liu and Gildea (2005), but did not achieve higher correlation with human *adequacy* judgments than metrics like METEOR. Another automatic metric, ULC (Giménez and Màrquez, 2007, 2008), incorporates several semantic similarity features and shows improved correlation with human judgement of translation quality (Callison-Burch *et al.*, 2007; Giménez and Màrquez, 2007; Callison-Burch *et al.*, 2008; Giménez and Màrquez, 2008) but no work has been done towards tuning an SMT system using a pure form of ULC perhaps due to its expensive run time. Likewise, SPEDE (Wang and Manning, 2012) predicts the edit sequence needed to match the machine translation to the reference translation via an integrated probabilistic FSM and probabilistic PDA model. The semantic textual similarity metric Sagan (Castillo and Estrella, 2012) is based on a complex textual entailment pipeline. These aggregated metrics require sophisticated feature extraction steps, contain many parameters that need to be tuned, and employ expensive linguistic resources such as WordNet or paraphrase tables. The expensive training, tuning and/or running time renders these metrics difficult to use in the SMT training cycle.

3 IMEANT

In this section we give a contrastive description of IMEANT: we first summarize the MEANT approach, and then explain how IMEANT differs.

3.1 Variants of MEANT

MEANT and its variants (Lo *et al.*, 2012) measure weighted f-scores over corresponding semantic frames and role fillers in the reference and machine translations. The automatic versions of MEANT replace humans with automatic SRL and alignment algorithms. MEANT typically outperforms BLEU, NIST, METEOR, WER, CDER and TER in correlation with human adequacy judgment, and is relatively easy to port to other languages, requiring only an automatic semantic parser and a monolingual corpus of the output language, which is used to gauge lexical similarity between the semantic role fillers of the reference and translation. MEANT is computed as follows:

1. Apply an automatic shallow semantic parser to both the reference and machine translations. (Figure 1 shows examples of automatic shallow semantic parses on both reference and MT.)

2. Apply the maximum weighted bipartite matching algorithm to align the semantic frames between the reference and machine translations according to the lexical similarities of the predicates. (Lo and Wu (2013a) proposed a backoff algorithm that evaluates the entire sentence of the MT output using the lexical similarity based on the context vector model, if the automatic shallow semantic parser fails to parse the reference or machine translations.)

3. For each pair of the aligned frames, apply the maximum weighted bipartite matching algorithm to align the arguments between the reference and MT output according to the lexical similarity of role fillers.

4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers according to the following definitions:

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG } j \text{ of aligned frame } i \text{ in MT} \\
q^1_{i,j} &\equiv \text{ARG } j \text{ of aligned frame } i \text{ in REF} \\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame } i \text{ of MT}}{\text{total \#tokens in MT}} \\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame } i \text{ of REF}}{\text{total \#tokens in REF}} \\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates} \\
w_j &\equiv \text{weight of similarity of ARG } j \\
\mathbf{e}_{i,\text{pred}} &\equiv \text{the pred string of the aligned frame } i \text{ of MT} \\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the pred string of the aligned frame } i \text{ of REF} \\
\mathbf{e}_{i,j} &\equiv \text{the role fillers of ARG } j \text{ of the aligned frame } i \text{ of MT} \\
\mathbf{f}_{i,j} &\equiv \text{the role fillers of ARG } j \text{ of the aligned frame } i \text{ of REF} \\
s(e,f) &= \text{lexical similarity of token } e \text{ and } f
\end{aligned}
$$

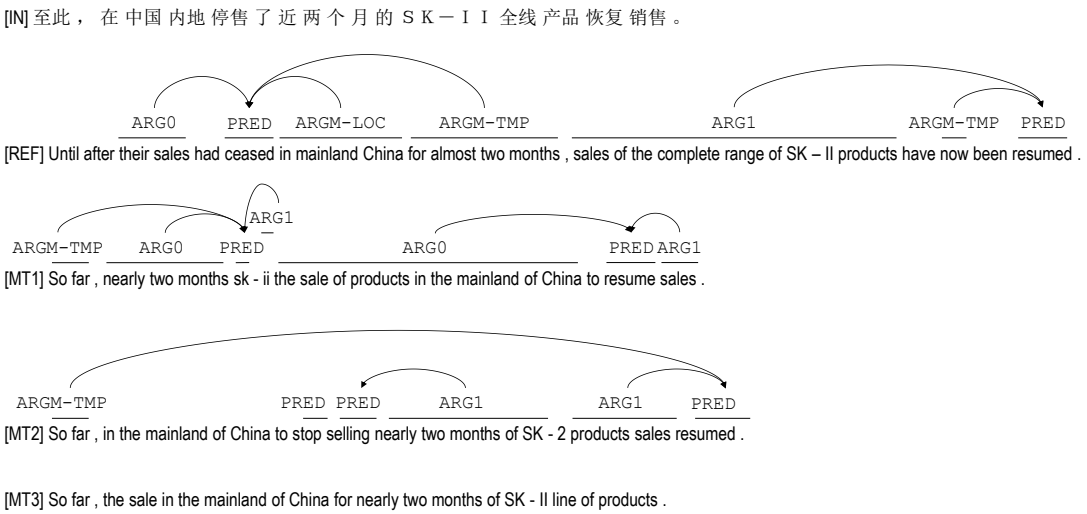

Figure 1: Examples of automatic shallow semantic parses. Both the reference and machine translations are parsed using automatic English SRL. There are no semantic frames for MT3 since there is no predicate in the MT output.

$$\text{prec}_{\mathbf{e},\mathbf{f}} = \frac{\sum_{e \in \mathbf{e}} \max_{f \in \mathbf{f}} s(e,f)}{|\,\mathbf{e}\,|}$$

$$\text{rec}_{\mathbf{e},\mathbf{f}} = \frac{\sum_{f \in \mathbf{f}} \max_{e \in \mathbf{e}} s(e,f)}{|\,\mathbf{f}\,|}$$

$$s_{i,\text{pred}} = \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}$$

$$s_{i,j} = \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}$$

$$\text{precision} = \frac{\sum_i w_i^0 \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q_{i,j}^0|}}{\sum_i w_i^0}$$

$$\text{recall} = \frac{\sum_i w_i^1 \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q_{i,j}^1|}}{\sum_i w_i^1}$$

$$\text{MEANT} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where $q_{i,j}^0$ and $q_{i,j}^1$ are the argument of type j in frame i in MT and REF respectively. w_i^0 and w_i^1 are the weights for frame i in MT/REF respectively. These weights estimate the degree of contribution of each frame to the overall meaning of the sentence. w_{pred} and w_j are the weights of the lexical similarities of the predicates and role fillers of the arguments of type j of all frame between the reference translations and the MT output.There is a total of 12 weights for the set of semantic role labels in MEANT as defined in Lo and Wu (2011b). For MEANT, they are determined using supervised estimation via a simple grid search to optimize the correlation with human adequacy judgments (Lo and Wu, 2011a). For UMEANT (Lo and Wu, 2012), they are estimated in an unsupervised manner using relative frequency of each semantic role label in the references and thus UMEANT is useful when human judgments on adequacy of the development set are unavailable.

$s_{i,\text{pred}}$ and $s_{i,j}$ are the lexical similarities based on a context vector model of the predicates and role fillers of the arguments of type j between the reference translations and the MT output. Lo *et al.* (2012) and Tumuluru *et al.* (2012) described how the lexical and phrasal similarities of the semantic role fillers are computed. A subsequent variant of the aggregation function inspired by Mihalcea *et al.* (2006) that normalizes phrasal similarities according to the phrase length more accurately was used in more recent work (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b). In this paper, we will assess IMEANT against the latest version of MEANT (Lo *et al.*, 2014) which, as shown, uses f-score to aggregate individual token similarities into the composite phrasal similarities of semantic role fillers,since this has been shown to be more accurate than the previously used aggregation functions.

Recent studies (Lo *et al.*, 2013a; Lo and Wu, 2013a; Lo *et al.*, 2013b) show that tuning MT systems against MEANT produces more robustly adequate translations than the common practice of tuning against BLEU or TER across different data genres, such as formal newswire text, informal web forum text and informal public speech.

In an alternative quality-estimation oriented line of research, Lo *et al.* (2014) describe a cross-lingual variant called XMEANT capable of evaluating translation quality without the need for expensive human reference translations, by utilizing semantic parses of the original foreign input sentence instead of a reference translation. Since XMEANT's results could have been due to either (1) more accurate evaluation of phrasal similarity via cross-lingual translation probabilities, or (2) better match of semantic frames without reference translations, there is no direct evidence whether ITGs contribute to the improvement in MEANT's correlation with human adequacy judgment. For the sake of better understanding whether ITGs improve semantic MT evaluation, we will also assess IMEANT against cross-lingual XMEANT.

3.2 The IMEANT metric

Although MEANT was previously shown to produce higher correlation with human adequacy judgments compared to other automatic metrics, our error analyses suggest that it still suffers from a common weakness among metrics employing lexical similarity, namely that word/token alignments between the reference and machine translations are severely under constrained. No bijectivity or permutation restrictions are applied, even between compositional segments where this should be natural. This can cause role fillers to be aligned even when they should not be. IMEANT, in contrast, uses a bracketing inversion transduction grammar to constrain permissible token alignment patterns between aligned role filler phrases. The semantic frames above the token level also fits ITG compositional structure, consistent with the aforementioned semantic frame alternation coverage study of Addanki *et al.* (2012). Figure 2 illustrates how the ITG constraints are consistent with the needed permutations between semantic role fillers across the reference and machine translations for a sample sentence from our evaluation data, which as we will see leads to higher HAJ correlations than MEANT.

Subject to the structural ITG constraints, IMEANT scores sentence translations in a spirit similar to the way MEANT scores them: it utilizes an aggregated score over the matched semantic role labels of the automatically aligned semantic frames and their role fillers between the reference and machine translations. Despite the structural differences, like MEANT, at the conceptual level IMEANT still aims to evaluate MT output in terms of the degree to which the translation has preserved the essential "who did what to whom,for whom, when, where, how and why" of the foreign input sentence.

Unlike MEANT, however, IMEANT aligns and scores under ITG assumptions. MEANT uses a maximum alignment algorithm to align the tokens in the role fillers between the reference and machine translations, and then scores by aggregating the lexical similarities into a phrasal similarity using an f-measure. In contrast, IMEANT aligns and scores by utilizing a length-normalized weighted BITG (Wu, 1997; Zens and Ney, 2003; Saers and Wu, 2009; Addanki *et al.*, 2012). To be precise in this regard, we can see IMEANT as differing from the foregoing description of MEANT in the definition of $s_{i,\text{pred}}$ and $s_{i,j}$, as follows.

$$
\begin{aligned}
G &\equiv \langle \{\mathrm{A}\}, \mathcal{W}^0, \mathcal{W}^1, \mathcal{R}, \mathrm{A} \rangle \\
\mathcal{R} &\equiv \{\mathrm{A} \rightarrow [\mathrm{AA}], \mathrm{A} \rightarrow \langle \mathrm{AA} \rangle, \mathrm{A} \rightarrow e/f\}
\end{aligned}
$$

$$
\begin{aligned}
p([\mathrm{AA}]\,|\mathrm{A}) &= p(\langle \mathrm{AA} \rangle | \mathrm{A}) = 1 \\
p(e/f|\mathrm{A}) &= s(e,f) \\
s_{i,\text{pred}} &= \lg^{-1}\left(\frac{\lg\left(P\left(\mathrm{A} \overset{*}{\Rightarrow} \mathbf{e}_{i,\text{pred}}/\mathbf{f}_{i,\text{pred}}|G\right)\right)}{\max(|\,\mathbf{e}_{i,\text{pred}}\,|,|\,\mathbf{f}_{i,\text{pred}}\,|)}\right) \\
s_{i,j} &= \lg^{-1}\left(\frac{\lg\left(P\left(\mathrm{A} \overset{*}{\Rightarrow} \mathbf{e}_{i,j}/\mathbf{f}_{i,j}|G\right)\right)}{\max(|\,\mathbf{e}_{i,j}\,|,|\,\mathbf{f}_{i,j}\,|)}\right)
\end{aligned}
$$

where G is a bracketing ITG whose only non terminal is A, and $\mathcal{R}$ is a set of transduction rules with $e \in \mathcal{W}^0 \cup \{\epsilon\}$ denoting a token in the MT output (or the *null* token) and $f \in \mathcal{W}^1 \cup \{\epsilon\}$ denoting a token in the reference translation (or the *null* token). The rule probability (or more accurately, rule weight) function p is set to be 1 for structural transduction rules, and for lexical transduction rules it is defined using MEANT's context vector model based lexical similarity measure. To calculate the inside probability (or more accurately, inside score) of a pair of segments, $P\left(\mathrm{A} \overset{*}{\Rightarrow} \mathbf{e}/\mathbf{f}|G\right)$, we use the algorithm described in Saers *et al.* (2009). Given this, $s_{i,\text{pred}}$ and $s_{i,j}$ now represent the length normalized BITG parse scores of the predicates and role fillers of the arguments of type j between the reference and machine translations.

4 Experiments

In this section we discuss experiments indicating that IMEANT further improves upon MEANT's

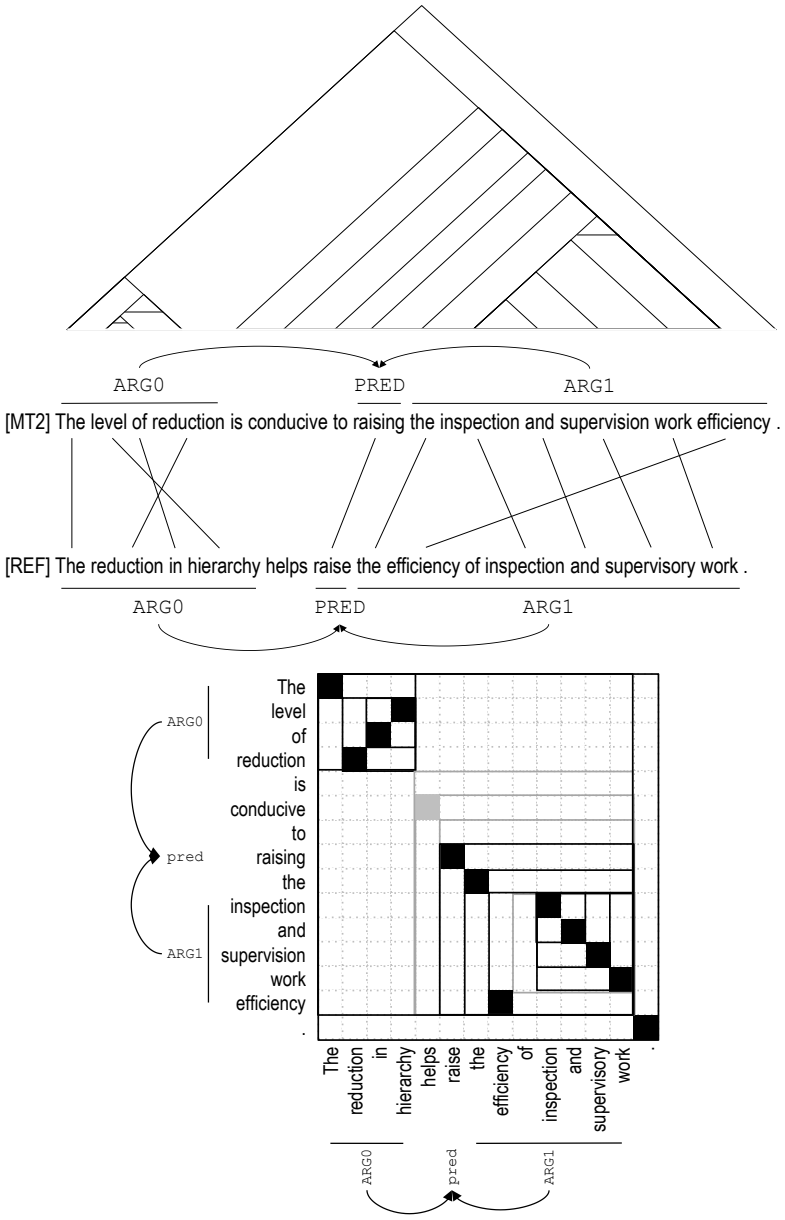

Figure 2: An example of aligning automatic shallow semantic parses under ITGs, visualized using both biparse tree and alignment matrix depictions, for the Chinese input sentence 层级的减少有利于提高检查监督工作的效率。. Both the reference and machine translations are parsed using automatic English SRL. Compositional alignments between the semantic frames and the tokens within role filler phrases obey inversion transduction grammars.

already-high correlation with human adequacy judgments.

4.1 Experimental setup

We perform the meta-evaluation upon two different partitions of the DARPA GALE P2.5 Chinese-English translation test set. The corpus includes the Chinese input sentences, each accompanied by one English reference translation and three participating state-of-the-art MT systems' output.

For the sake of consistent comparison, the first evaluation partition, GALE-A, is the same as the one used in Lo and Wu (2011a), and the second evaluation partition, GALE-B, is the same as the one used in Lo and Wu (2011b).

For both reference and machine translations, the ASSERT (Pradhan *et al.*, 2004) semantic role labeler was used to automatically predict semantic parses.

Table 1: Sentence-level correlation with human adequacy judgements on different partitions of GALE P2.5 data. IMEANT always yields top correlations, and is more consistent than either MEANT or its recent cross-lingual XMEANT quality estimation variant. For reference, the human HMEANT upper bound is 0.53 for GALE-A and 0.37 for GALE-B—thus, the fully automated IMEANT approximation is not far from closing the gap.

metric	*GALE-A*	*GALE-B*
IMEANT	**0.51**	**0.33**
XMEANT	**0.51**	0.20
MEANT	0.48	**0.33**
METEOR 1.5 (2014)	0.43	0.10
NIST	0.29	0.16
METEOR 0.4.3 (2005)	0.20	0.29
BLEU	0.20	0.27
TER	0.20	0.19
PER	0.20	0.18
CDER	0.12	0.16
WER	0.10	0.26

4.2 Results

The sentence-level correlations in Table 1 show that IMEANT outperforms other automatic metrics in correlation with human adequacy judgment. Note that this was achieved with no tuning whatsoever of the default rule weights (suggesting that the performance of IMEANT could be further improved in the future by slightly optimizing the ITG weights).

On the GALE-A partition, IMEANT shows 3 points improvement over MEANT, and is tied with the cross-lingual XMEANT quality estimator discussed earlier.IMEANT produces much higher HAJ correlations than any of the other metrics.

On the GALE-B partition, IMEANT is tied with MEANT, and is significantly better correlated with HAJ than the XMEANT quality estimator. Again, IMEANT produces much higher HAJ correlations than any of the other metrics.

We note that we have also observed this pattern consistently in smaller-scale experiments—while the monolingual MEANT metric and its cross-lingual XMEANT cousin vie with each other on different data sets, IMEANT robustly and consistently produces top HAJ correlations.

In both the GALE-A and GALE-B partitions, IMEANT comes within a few points of the human upper bound benchmark HAJ correlations computed using the human labeled semantic frames and alignments used in the HMEANT.

Data analysis reveals two reasons that IMEANT correlates with human adequacy judgement more closely than MEANT. First, BITG constraints indeed provide more accurate phrasal similarity aggregation, compared to the naive bag-of-words based heuristics employed in MEANT. Similar results have been observed while trying to estimate word alignment probabilities where BITG constraints outperformed alignments from GIZA++ (Saers and Wu, 2009).

Secondly, the permutation and bijectivity constraints enforced by the ITG provide better leverage to reject token alignments when they are not appropriate, compared with the maximal alignment approach which tends to be rather promiscuous. A case of this can be seen in Figure 3, which shows the result on the same example sentence as in Figure 1. Disregarding the semantic parsing errors arising from the current limitations of automatic SRL tools, the ITG tends to provide clean, sparse alignments for role fillers like the ARG1 of the resumed PRED, preferring to leave tokens like complete and range unaligned instead of aligning them anyway as MEANT's maximal alignment algorithm tends to do. Note that it is not simply a matter of lowering thresholds for accepting token alignments: Tumuluru *et al.* (2012) showed that the competitive linking approach (Melamed, 1996) which also generally produces sparser alignments does not work as well in MEANT, whereas the ITG appears to be selective about the token alignments in a manner that better fits the semantic structure.

For contrast, Figure 4 shows a case where IMEANT appropriately accepts dense alignments.

5 Conclusion

We have presented IMEANT, an inversion transduction grammar based rethinking of the MEANT semantic frame based MT evaluation approach, that achieves higher correlation with human adequacy judgments of MT output quality than MEANT and its variants, as well as other common evaluation metrics. Our results improve upon previous research showing that MEANT's explicit use of semantic frames leads to state-of-the-art automatic MT evaluation. IMEANT achieves this by aligning and scoring semantic frames under a simple, consistent ITG that provides empirically

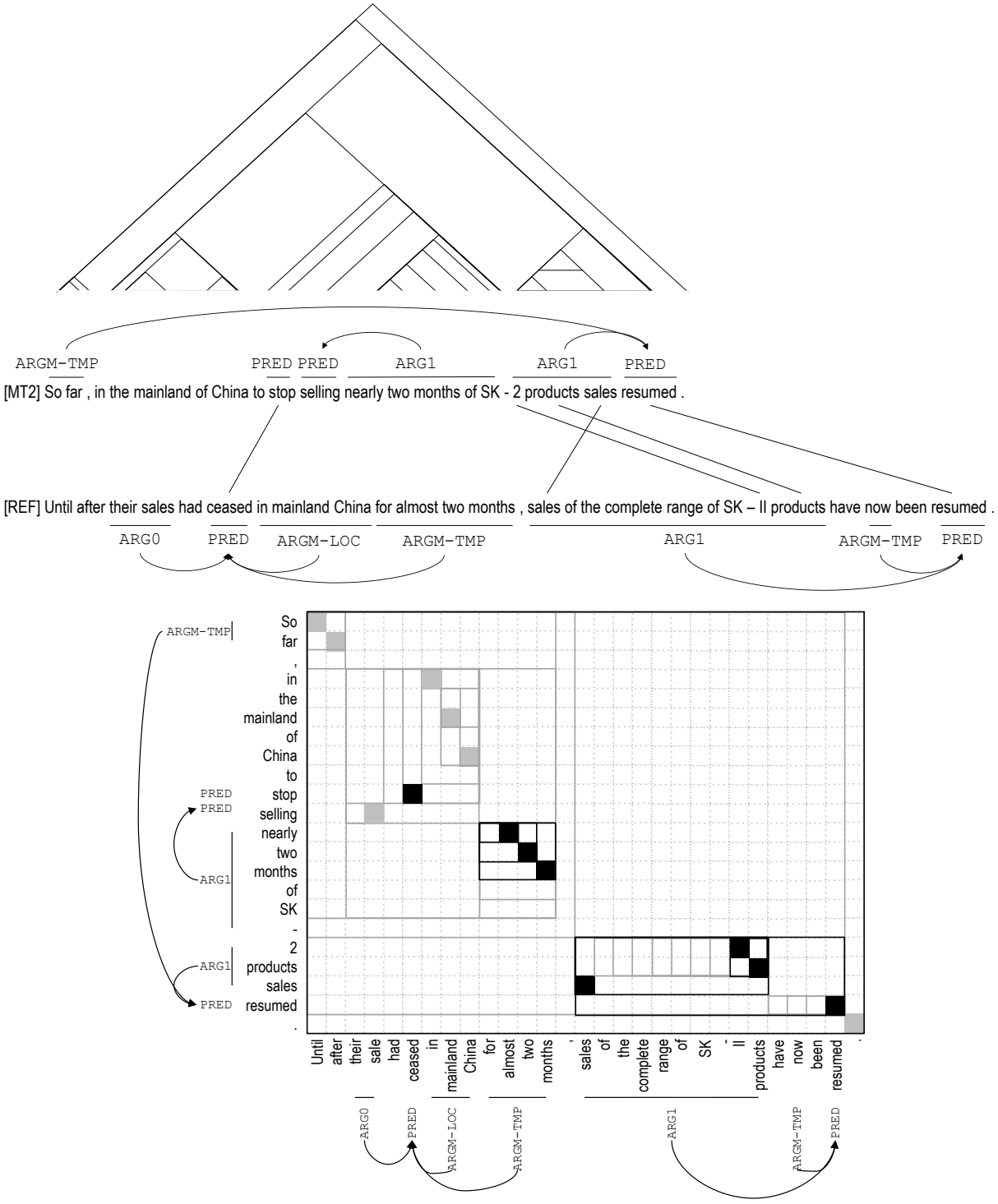

Figure 3: An example where the ITG helps produce correctly sparse alignments by rejecting inappropriate token alignments in the ARG1 of the resumed PRED, instead of wrongly aligning tokens like the, complete, and range as MEANT tends to do. (The semantic parse errors are due to limitations of automatic SRL.)

informative permutation and bijectivity biases, instead of the maximal alignment and bag-of-words assumptions used by MEANT. At the same time, IMEANT retains the Occam's Razor style simplicity and representational transparency characteristics of MEANT.

Given the absence of any tuning of ITG weights in this first version of IMEANT, we speculate that

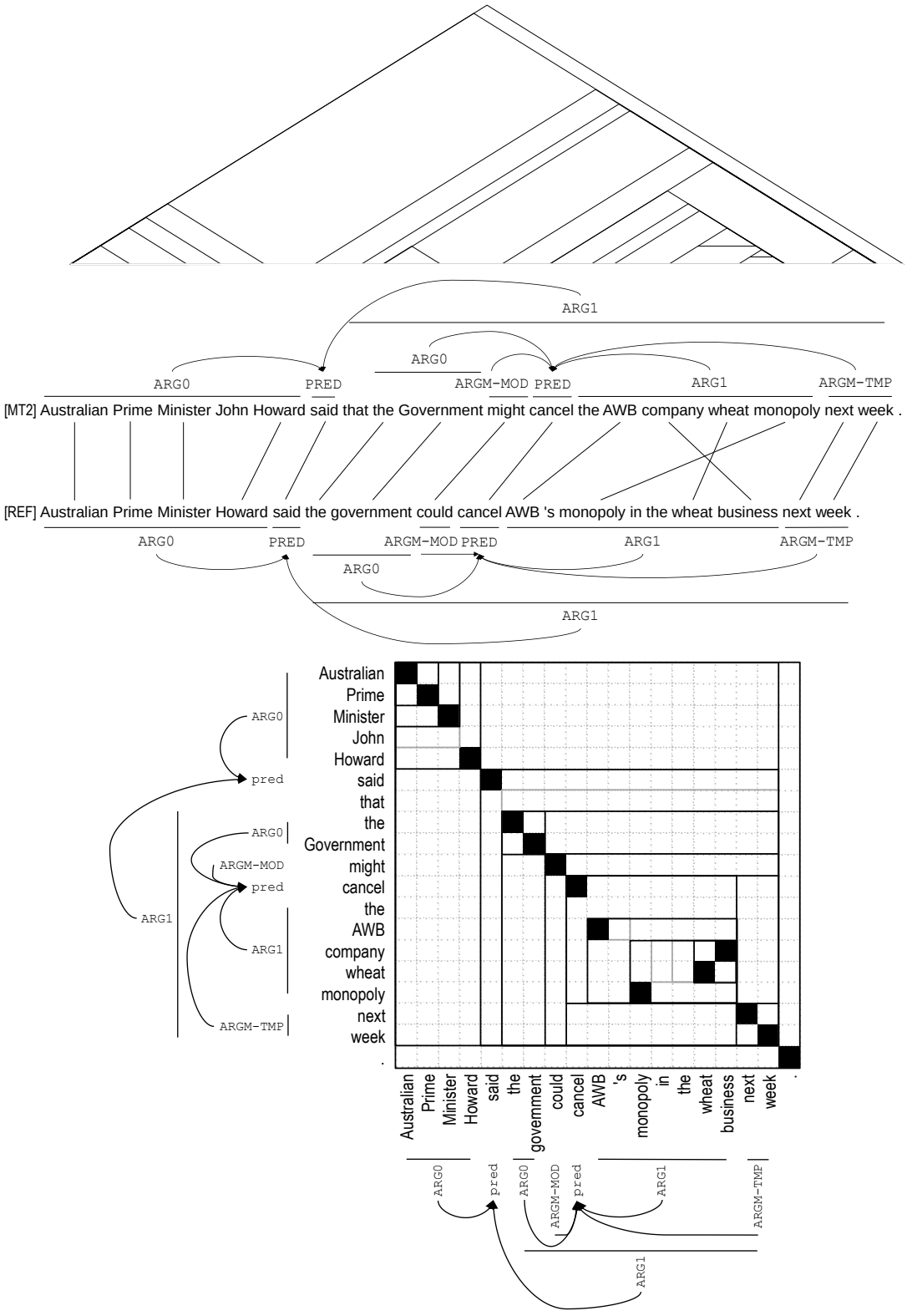

Figure 4: An example of dense alignments in IMEANT, for the Chinese input sentence 澳大利亚总理霍华德表示，政府可能于下周取消 AWB 公司小麦专卖的业务。 (The semantic parse errors are due to limitations of automatic SRL.)

IMEANT could perform even better than it already does here.We plan to investigate simple hyperparameter optimizations in the near future.

6 Acknowledgments

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) under BOLT contract nos. HR0011-12-C-0014 and 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. Thanks to Karteek Addanki for supporting work, and to Pascale Fung, Yongsheng Yang and Zhaojun Wu for sharing the maximum entropy Chinese segmenter and C-ASSERT, the Chinese semantic parser.

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.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluating the role of BLEU in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, 2006.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (meta-) evaluation of machine translation. In *Second Workshop on Statistical Machine Translation (WMT-07)*, 2007.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Third Workshop on Statistical Machine Translation (WMT-08)*, 2008.

Julio Castillo and Paula Estrella. Semantic textual similarity for MT evaluation. In *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

Michael Denkowski and Alon Lavie. METEOR universal: Language specific translation evaluation for any target language. In *9th Workshop on Statistical Machine Translation (WMT 2014)*, 2014.

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.

Jesús Giménez and Lluís Màrquez. Linguistic features for automatic evaluation of heterogenous MT systems. In *Second Workshop on Statistical Machine Translation (WMT-07)*, pages 256–264, Prague, Czech Republic, June 2007.

Jesús Giménez and Lluís Màrquez. A smorgasbord of features for automatic MT evaluation. In *Third Workshop on Statistical Machine Translation (WMT-08)*, Columbus, Ohio, June 2008.

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Workshop on Statistical Machine Translation (WMT-06)*, 2006.

Gregor Leusch and Hermann Ney. Bleusp, invwer, cder: Three improved mt evaluation measures. In *NIST Metrics for Machine Translation Challenge (MetricsMATR), at Eighth Conference of the Association for Machine Translation in the Americas (AMTA 2008)*, Waikiki, Hawaii, Oct 2008.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. A novel string-to-string distance measure with applications to machine translation evaluation. In *Machine Translation Summit IX (MT Summit IX)*, New Orleans, Sep 2003.

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.

Ding Liu and Daniel Gildea. Syntactic features for evaluation of machine translation. In *Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, June 2005.

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. SMT vs. AI redux: How semantic frames evaluate MT more accurately. In *Twenty-second International Joint Conference on Artificial Intelligence (IJCAI-11)*, 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.

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.

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.

I. Dan Melamed. Automatic construction of clean broad-coverage translation lexicons. In *2nd Conference of the Association for Machine Translation in the Americas (AMTA-1996)*, 1996.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *The Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, volume 21, 2006.

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.

Karolina Owczarzak, Josef van Genabith, and Andy Way. Dependency-based automatic evaluation for machine translation. In *Syntax and Structure in Statistical Translation (SSST)*, 2007.

Karolina Owczarzak, Josef van Genabith, and Andy Way. Evaluating machine translation with LFG dependencies. *Machine Translation*, 21:95–119, 2007.

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.

Miguel Rios, Wilker Aziz, and Lucia Specia. TINE: A metric to assess MT adequacy. In *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, 2011.

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.

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.

Anand Karthik Tumuluru, Chi-kiu Lo, and Dekai Wu. Accuracy and robustness in measuring the lexical similarity of semantic role fillers for automatic semantic MT evaluation. In *26th Pacific Asia Conference on Language, Information, and Computation (PACLIC 26)*, 2012.

Mengqiu Wang and Christopher D. Manning. SPEDE: Probabilistic edit distance metrics for MT evaluation. In *7th Workshop on Statistical Machine Translation (WMT 2012)*, 2012.

Dekai Wu. An algorithm for simultaneously bracketing parallel texts by aligning words. In *33rd Annual Meeting of the Association for Computational Linguistics (ACL 95)*, pages 244–251, Cambridge, Massachusetts, June 1995.

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.