EUROPEAN ASSOCIATION FOR MACHINE TRANSLATION

EAMT2018

Proceedings of the

21st Annual Conference of the European Association for Machine Translation

28–30 May 2018

Universitat d'Alacant

Alacant, Spain

Edited by

Juan Antonio Pérez-Ortiz

Felipe Sánchez-Martínez

Miquel Esplà-Gomis

Maja Popović

Celia Rico

André Martins

Joachim Van den Bogaert

Mikel L. Forcada

Organised by

Universitat d'Alacant

Universidad de Alicante

transducens

research group

The papers published in this proceedings are —unless indicated otherwise— covered by the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 International (CC-BY-ND 3.0). You may copy, distribute, and transmit the work, provided that you attribute it (authorship, proceedings, publisher) in the manner specified by the author(s) or licensor(s), and that you do not use it for commercial purposes. The full text of the licence may be found at https://creativecommons.org/licenses/by-nc-nd/3.0/deed.en.

© 2018 The authors
ISBN: 978-84-09-01901-4

SRL for low resource languages isn't needed for semantic SMT

Meriem Beloucif and **Dekai Wu**
Human Language Technology Center
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
mbeloucif|dekai@cs.ust.hk

Abstract

Previous attempts at injecting semantic frame biases into SMT training for low resource languages failed because either (a) no semantic parser is available for the low resource input language; or (b) the output English language semantic parses excise relevant parts of the alignment space too aggressively. We present the first semantic SMT model to succeed in significantly improving translation quality across many low resource input languages for which no automatic SRL is available —consistently and across all common MT metrics. The results we report are the best by far to date for this type of approach; our analyses suggest that in general, easier approaches toward including semantics in training SMT models may be more feasible than generally assumed even for low resource languages where semantic parsers remain scarce.

While recent proposals to use the crosslingual evaluation metric XMEANT during inversion transduction grammar (ITG) induction are inapplicable to low resource languages that lack semantic parsers, we break the bottleneck via a vastly improved method of biasing ITG induction toward learning more semantically correct alignments using the *monolingual* semantic evaluation metric MEANT. Unlike XMEANT, MEANT requires only a readily-available English (output language) semantic parser. The advances we report here exploit the novel realization that MEANT represents an excellent way to semantically bias expectation-maximization induction even for low resource languages. We test our systems on challenging languages including Amharic, Uyghur, Tigrinya and Oromo. Results show that our model influences the learning towards more semantically correct alignments, leading to better translation quality than both the standard ITG or GIZA++ based SMT training models on different datasets.

1 Introduction

Statistical machine translation (SMT) for low resource languages has been a difficult task due to the unavailability of large parallel corpora. It becomes imperative to make learning from *small* data more efficient by adding additional constraints to create stronger inductive biases—especially linguistically well-motivated constraints, such as the shallow semantic parses of the training sentences. However, while automatic semantic role labeling (SRL) is readily available to produce shallow semantic parses for a high-resource *output* language (typically English), the problem is that SRL is usually not available for low resource *input* languages such as Tigrinya, Oromo, Uyghur or Uzbek.

In this paper, we propose a new method which adopts the monolingual semantic evaluation metric MEANT as a confidence-weighting measure to assess the degree of goodness of training instances, giving a newer strategy than Beloucif and Wu (2016a) who used the degree of compatibility or similarity between the semantic role labeling of the input and output sentences. Their approach might outperform ours for high-resource languages, but is completely inapplicable to low resource languages because XMEANT requires both the input and output semantic parses – whereas MEANT does not require an SRL parse for the low resource input language.

© 2018 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 59–68
Alacant, Spain, May 2018.

Additionally, we also introduce a notion of **semantic role labeling coverage** as a second English monolingual confidence-weighting measure. An SRL coverage score roughly quantifies what proportion of a sentence is accounted for by a shallow semantic parse. The variety of approaches proposed here belong to a family of semantic SMT methods that has recently been advanced, wherein SRL constraints or biases are injected very early in the SMT training pipeline so as to maximize their influence on what translation model is learned. We test our models on multiple difficult low resource translation tasks: Amharic, Somali, Tigrinya, Oromo, Uzbek and Uyghur always translating into English. Despite having SRLs only on the English side, we show that our models influence the learning toward more semantically correct alignments. Our results show that this way of inducing ITGs gives a better translation quality than the conventional ITG (Saers and Wu, 2009) and the traditional GIZA++ (Och and Ney, 2000) alignments.

2 Related work

2.1 Semantic frames in the SMT pipeline

Semantic role labeling (SRL) or shallow semantic parsing, is a task that defines the semantic event structure *who did what to whom*, *for whom*, *when*, *where*, *how* and *why* in a given sentence (Gildea and Jurafsky, 2002). Only a few works integrate information provided by an SRL in SMT. However, most of the approaches do not use SRL for training, but either for tuning, evaluation or post-processing. For instance, Wu and Fung (2009) have empirically shown that including SRL for post-processing the MT output improves the translation quality. Their method maximizes the crosslingual match of the semantic labels between the input and the output sentences. Many tools that use SRL for MT evaluation have been proposed such as the semantic evaluation metric MEANT, which adopts the principle that a good translation preserves the semantic event structure across translations (Lo and Wu, 2011a, 2012; Lo *et al.*, 2012) or XMEANT (Lo *et al.*, 2014), the crosslingual version of MEANT, which uses the foreign input instead of the reference translation.

Liu and Gildea (2010) and Aziz *et al.* (2011) use input language SRL to train a tree-to-string SMT system. Xiong *et al.* (2012) trained a two pass discriminative model to incorporate source side predicate-argument structures into SMT. Komachi *et al.* (2006) and Wu *et al.* (2011) preprocess the input sentence to match the verb frame alternations in the output side. Moreover, Beloucif *et al.* (2015) have shown that including a semantic frame based objective function at an early stage of training SMT systems gives better translations than relying on tuning loglinear weights against a semantic based objective function such as MEANT. All these approaches are *inapplicable* when translating low resource languages since they either require the input language semantic parse or both languages SRL parses.

The most recent work that includes SRL during the actual learning of bilingual constituents for *low resource languages* is the one by Beloucif and Wu (2016b). However, our approach is quite different in spirit, and significantly outperforms theirs. Whereas their method for training ITGs penalizes bilingual constituents in the expectation-maximization (EM) biparse forests when they violate an English SRL, our training approach weights entire bilingual sentence pairs by predicting a confidence derived from MEANT. The problem with their approach is that they attempt to demote some partial hypotheses during the ITG training, which can excise relevant parts of the alignment search space aggressively.

2.2 The semantic based evaluation metric MEANT

The main model we propose adopts MEANT (Lo and Wu, 2011a, 2012; Lo *et al.*, 2012) to confidence-weight training instances. MEANT is a semantic frame based evaluation metric which compares the SRL parse of the MT output against the SRL parse of the reference translations provided. Then it produces a score that assesses the degree of similarity between their semantic frame structures. The MEANT algorithm is described in figure 1.

In figure 1, $q^0_{i,j}$ and $q^1_{i,j}$ are the arguments of type j in frame i in MT and REF respectively. w^0_i and w^1_i are the weights for frame i in MT/REF respectively.

The weights mentioned in the algorithm 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

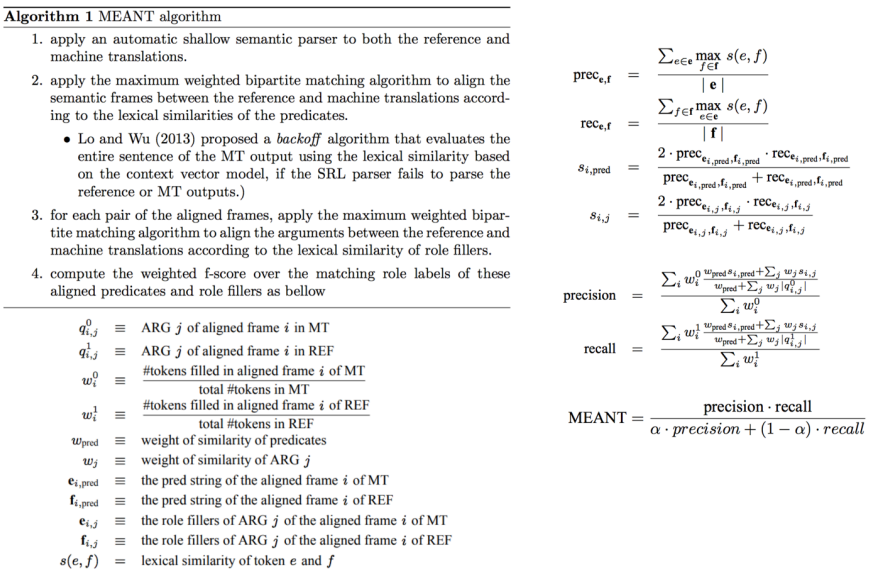

Figure 1: The MEANT algorithm from left to right.

j of all frame between the reference translations and the machine translations. There is a total of 12 weights for the set of semantic role labels in MEANT as defined in Lo and Wu (2011b). They are determined using supervised estimation via a simple grid search to optimize the correlation with human adequacy judgments (Lo and Wu, 2011a).

3 Core model

The approaches proposed in this work inject a form of semantic parse bias into early stage word alignment using ITG (Wu, 1997) training, which (as shown in the results section) outperforms conventional GIZA++ (Och and Ney, 2000) based intersection/union-of-bidirectional-IBM-word-alignment strategies. Specifically, our defined approaches assume a token based BITG (bracketing ITG) (Wu, 1997) system, a choice based on previous works showing that: (a) BITG based alignments outperform GIZA++ alignments (Saers *et al.*, 2009); (b) ITG alignments have been empirically shown to cover almost 100% of *semantic* frame alternations, while ruling out the majority of incorrect alignments (Addanki *et al.*, 2012). The BITG model used in this work is initialized with uniform structural probabilities, setting aside half of the probability mass for lexical rules. The lexical 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 EM, where the expectation step is calculated using beam pruned parsing (Saers *et al.*, 2009) with a beam width of 100. In the last iteration, the alignments imposed by the Viterbi parses are extracted as the final word alignments.

Saers and Wu (2011) showed how to compute expectations for EM re-estimation with outside probabilities as follows:

$$E_\theta = \frac{\alpha(M \to AL)\beta(M \to AL)}{\alpha(S_{0,|e|,0,|f|})\beta(S_{0,|e|,0,|f|})} \quad (1)$$

where $\alpha(M \to AL)$ and $\beta(M \to AL)$ are the inside and the outside probabilities of the derivation $M \to AL$ respectively. $\alpha(S_{0,|e|,0,|f|})$ is the initial inside probability, while $\beta(S_{0,|e|,0,|f|})$ represents the initial outside probability. Traditionally, the outside probability $\beta(S_{0,|e|,0,|f|})$ in the inside-outside algorithm is set to 1.0 as it represents the number of observations of a training instance (each bisentence is observed once). An intuitive way to distinguish good from bad sentences would be to favor sentences that have a good semantic parse, by setting the outside probability to be a weight (a fractional count between 0 and 1) that somehow reflects the goodness of the semantic parse better than a unified fractional count. Therefore, biasing the learning towards training instances which have a good SRL parse.

4 MEANT as a training objective function

4.1 Injecting MEANT

A more robust way to assess the degree of goodness of training instances has been shown to be the crosslingual evaluation metric XMEANT Beloucif and Wu (2016a). Unfortunately, this is not applicable in low resource settings since XMEANT assesses the compatibility between the English output and the input foreign language—for which the semantic parse is unavailable. Instead of computing the crosslingual compatibility between the input and the output semantic parses, we adopt the monolingual semantic frame evaluation metric MEANT as a confidence measure.

The evaluation metric MEANT computes the semantic frame coverage between the input and the MT reference. We propose to use MEANT as a confidence-weight measure by computing the semantic frame coverage in the English sentence. We obtain the SRL coverage of a sentence by computing the MEANT score between the input English sentence and the same sentence as a reference. We do not take into account the chunks that have no semantic parse (*backoff* was mentioned in figure 2).

Figure 2 illustrates two out of three possible situations for applying MEANT as a confidence-weight measure. The sentences that are fully semantically parsed like [ARG0 I][TARGET ate][ARG1 an apple]. have a MEANT score equal to 1.0. If the sentence is partially SRLed, the MEANT score is less than 1.0. For instance, the MEANT score for the parse Where do [ARG0 I][TARGET get][ARG2 off] to go to Union Square? is less than 1, but higher than 0. Furthermore, we note that a few sentences have a 0 MEANT score. In fact, we have experimented with three automatic SRLs: ASSERT (Pradhan *et al.*, 2004), MATE (Björkelund *et al.*, 2009) and MATEPLUS (Roth and Woodsend, 2014); we have observed that these SRL systems completely fail to parse sentences containing the verb to be; sentences like the light was red are ignored. However, we show that even while ignoring sentences containing to be, our systems are still outperforming conventional models on multiple challenging low resource languages.

4.2 Injecting monolingual SRL coverage

The second new strategy for judging the reliability of training instances using semantics is the monolingual SRL coverage, which looks at the proportion of a sentence that is accounted for by the English semantic parse. In its simplest, monolingual form, we define the monolingual coverage as follows:

$$\varphi_1 = (\#\text{ labels} / \#\text{ words labelled}) + \beta_0 \qquad (2)$$

where β_0 is a hyperparameter that is manually set to avoid eliminating sentences with 0 probability. The intuition in this approach is to give a higher SRL coverage to sentences that are easily SRLed and a low coverage to complex sentences that are hard to parse by an automatic SRL. For instance, the SRL parse: okay, sure. [TARGET pay][ARG1 this] up front when you are ready. take your time would have a low coverage. These are the kind of sentences that we do not want to rely on during the training. This sentence is hard to semantically parse automatically and it is a bit colloquial which makes it a less favorable training instance, especially in a low resource setting where good training instances are hard to obtain. We have also experimented with another version of the coverage, which computes the coverage over the number of all the words instead of all the words that were labelled. The version described in equation (3) slightly outperforms the second model, thus we only report the former.

4.3 Injecting sentence length

The purpose of our experiments is to show that injecting a monolingual semantic based objective function for deriving ITG induction helps learn more semantically correct bilingual correlations. We propose an intuitive approach to evaluate the degree of goodness of sentence pairs based on the sentence length of the English side.

This method simply counts the number of words in a sentence; we then take the reverse sentence length as a confidence-weight. We claim that having long sentences makes the data more sparse when we train on a small corpus. This might prevent the system from efficiently learning from the data and thus hurts the translation quality. The reverse sentence length is calculated as follows:

$$\mathrm{L} = (1 / \#\text{ words}) \qquad (3)$$

We experiment this method with the Chinese–English translation task. We show in table 1 that using reverse sentence length as a confidence-weighting measure slightly improves the SMT

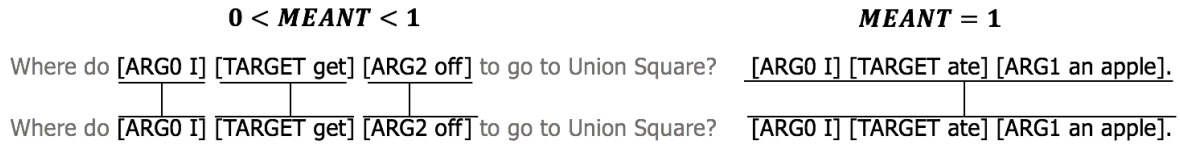

Figure 2: MEANT score in different situations.

Table 1: The monolingual SRL coverage model greatly outperforms the sentence length one.

Alignments	BLEU	TER
GIZA++	19.23	63.40
BITG	20.05	63.19
+Sentence length	20.54	62.49
+SRL_{en}	**23.60**	**61.68**

quality in terms of BLEU and TER scores in comparison to GIZA++ and BITG based models. This shows that confidence-weighting the training instances even with a simple measure like sentence length helps improve SMT for low resource languages. However, we note that our monolingual SRL coverage based model substantially improves the translation quality compared to using a simple heuristic such as sentence length.

5 Experimental setup

5.1 Training data

Our experiments aim to show that adopting MEANT as a semantic objective function to bias ITG induction at an early stage the SMT models' training helps reduce the need of extremely large corpora as typically used in SMT training. We focus on the generalization from *only* low resource data and thus focus our work on unpreprocessed data.

Table 2 represents the size of all datasets used in our experimental setup. Except for Chinese and Latvian, which are from IWSLT07 data and Europarl data Koehn (2005) respectively, all the other datasets are from the DARPA LORELEI program. The LORELEI data is diverse; it is composed of forums data and some Quranic verses. The IWSLT07 data is mainly spoken language. The size of the training data varies between 2K (Oromo) and 630K (Latvian) bisentences.

We purposely experiment with different language families including Turkic, Afro-asiatic, Indo-European and Sino-Tibetan languages to show that our approach is not language dependent and can easily be generalized across different languages. We deliberately experiment on a relatively small corpus for the two high-resource languages Chinese and Turkish; all the other languages are considered as low resource languages.

5.2 SMT pipeline

We test the different *alignments* described above using the standard MOSES toolkit (Koehn *et al.*, 2007), and a 6-gram language model learned with the SRI language model toolkit (Stolcke, 2002) trained on the English side of 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 and MIRA (Chiang, 2012).

5.3 NMT pipeline

Neural machine translation or NMT has been considered as a hot topic in machine translation over the past few years. NMT is a new encoder-decoder architecture for getting machines to learn to translate based on neural networks. Despite being relatively new, NMT has already shown promising results, achieving state-of-the-art performance for various language pairs (Luong and Manning, 2015; Sennrich *et al.*, 2015; Luong and Manning, 2016). For the sake of comparison, we set up a simple NMT baseline based on Neubig's toolkit lamtram (Neubig, 2015).

5.4 Tuning the hyperparameter for the monolingual SRL coverage based model

For the monolingual SRL coverage model, we tune the hyperparameter β_0 on Uzbek–English and Uyghur–English to find the best value of β_0. We test the model with the obtained hyperparameter with different language pairs. The tuning results are reported in table 3; although the difference in the results between the different values of β_0 is insignificant, we note that β_0=0.7 gives the best results across both language pairs. Therefore, we set

Table 2: The size of the different datasets in sentence pairs (foreign–English).

	Amharic	Chinese	Oromo	Somali	Tigrinya	Turkish	Uyghur	Uzbek	Latvian
Training	60,300	39,953	2,308	50,194	13,807	180,578	97,367	153,408	637,599
Tuning	3,016	1,512	116	2,510	691	1,000	2,000	1,200	2,000
Testing	3,015	489	116	2,510	691	500	1,000	600	2,000

Table 3: Tuning β_0 for the SRL coverage model.

	Uzbek–English		Uyghur–English	
Alignments	BLEU	TER	BLEU	TER
+SRL_{en} 1, β_0=0	18.29	**74.01**	23.67	66.02
+SRL_{en} 1, β_0=0.1	18.14	74.16	23.12	66.42
+SRL_{en} 1, β_0=0.5	18.11	74.18	23.70	65.74
+SRL_{en} 1, β_0=0.7	18.24	**74.03**	**23.85**	**65.57**
+SRL_{en} 1, β_0=1	**18.32**	74.56	23.43	66.75

β_0 to 0.7 in the remaining parts of the paper.

6 Results

Adopting MEANT for confidence-weighting gives the best results for translating low resource languages. We compare the performance of the MEANT and the monolingual English SRL coverage based BITG alignments against the conventional BITG and the traditional GIZA++ alignments. To efficiently assess the quality of our different systems, we evaluate using surface based metrics such as BLEU (Papineni *et al.*, 2002), edit-distance based metrics such as CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), PER (Tillmann *et al.*, 1997), TER (Snover *et al.*, 2006) and the semantic evaluation metric MEANT (Lo *et al.*, 2012).

6.1 Adopting MEANT gives the best results across multiple challenging low resource languages

Our experiments show that injecting the monolingual semantic evaluation metric MEANT as a training objective function gives the best results compared to any monolingual confidence-weighting model proposed so far since it consistently improves the translation quality for multiple challenging low resource languages. This can be explained by the fact that XMEANT and MEANT have the same constraints and thus we expect them to have the same behavior.

We note from table 4 that the alignments based on our proposed models (SRL_{en} is the monolingual SRL coverage and SRL_{MEANT} is the MEANT based model) achieve a much higher performance than the traditional GIZA++ and the unbiased BITG baseline across all metrics. The impact of MEANT or SRL coverage on the translation quality depends on the data size and on the nature of the language. Translation tasks like Oromo–English have harsher conditions than the Turkish–English task since Oromo data is harder to obtain. The highest scores that we managed to obtain on Oromo–English are 8.26 for BLEU and 11.33 for MEANT, which reflects the difficulty of the task we study here. In most cases, the difference varies between 2 BLEU points like in Amharic and Uzbek translations to 5 BLEU points like in the Chinese–English translation task. One exception is the Somali–English translation where we only note a small improvement (0.5 BLEU points); the reason is that the test data is too large (2500 sentences) in proportion to the size of the training data. Our methods seem to have a higher impact on error-rate metrics; we improved by around 13 PER points and 6 WER points on the Amharic–English translation task. We also improved semantic SMT by obtaining better MEANT scores on all our SRL based models.

However, the difference between the SRL coverage and the MEANT based models is small. The MEANT based model is better most of the time except for the Uzbek–English translation task, where the SRL coverage model is slightly better in terms of BLEU and TER.

Table 4: Adopting MEANT as a confidence-weighting measure produces the best results across all commonly used metrics.

	Amharic–English					
	MEANT	BLEU	TER	WER	PER	CDER
GIZA++	10.85	11.68	101.85	103.08	90.18	93.72
BITG	10.92	13.00	98.27	101.82	88.10	93.63
+ SRL_{en}	11.57	13.59	98.00	100.31	87.55	92.37
+ SRL_{MEANT}	**12.28**	**14.72**	**92.12**	**94.44**	**77.55**	**86.40**
	Chinese–English					
GIZA++	22.77	19.23	63.40	62.08	55.75	59.79
BITG	23.90	20.05	63.19	61.63	54.07	59.61
+ SRL_{en}	23.99	23.60	61.68	61.90	54.40	59.40
+ SRL_{MEANT}	**24.10**	**24.94**	**60.96**	**61.50**	**54.40**	**59.41**
	Uzbek–English					
GIZA++	14.47	17.09	80.91	87.71	64.61	78.11
BITG	16.55	17.66	78.12	84.60	62.86	75.51
+ SRL_{en}	17.04	**19.07**	**72.56**	78.99	57.34	70.36
SRL_{MEANT}	**17.35**	18.24	74.03	**78.63**	**57.00**	**70.00**
	Oromo–English					
GIZA++	9.59	5.16	134	134	110	124
BITG	10.04	7.80	131	131	113	121
+ SRL_{en}	10.40	7.92	126	129	111	122
SRL_{MEANT}	**11.33**	**8.26**	**123**	**125**	**105**	**119**
	Somali–English					
GIZA++	18.25	19.80	69.00	79.60	56.91	67.66
BITG	18.47	19.85	68.80	79.00	56.72	66.23
+ SRL_{en}	18.59	**20.24**	68.70	78.04	56.62	66.50
SRL_{MEANT}	**18.87**	20.06	**68.50**	**78.00**	**56.42**	**66.20**
	Tigrinya–English					
GIZA++	12.39	11.52	98.44	93.11	77.14	86.43
BITG	14.10	11.75	99.06	93.17	77.19	86.40
+ SRL_{en}	14.90	12.28	94.87	94.49	77.70	87.73
SRL_{MEANT}	**14.93**	**12.85**	**93.52**	**92.94**	**76.50**	**85.90**
	Turkish–English					
GIZA++	14.37	12.72	74.63	81.36	55.86	72.23
BITG	16.24	14.12	74.92	82.23	55.59	72.37
+ SRL_{en}	16.80	14.50	74.50	80.97	**53.78**	70.82
SRL_{MEANT}	**17.62**	**14.95**	**73.12**	**80.83**	54.12	**70.63**

Table 5: NMT models perform worse than SMT models for the Tigrinya–English translation task.

	BLEU	TER
SMT	11.52	98.44
SMT + SRL_{MEANT}	**12.85**	**94.87**
NMT	1.51	118
NMT + SRL_{MEANT}	1.91	99.16

6.2 NMT models are weak when translating low resource languages

Our goal is to investigate apples-to-apples comparison: (a) ability to generalize from *only* low resource data *without* transfer from related high-resource languages, and (b) ability to work with unpreprocessed data. We ran a simple NMT baseline with low resource languages. Neural NLP models in general and neural machine translation models in particular tend to need huge data to work

Table 6: MEANT based models perform well ina high resource setting, but the impact is higher in a low resource setting.

	MEANT	BLEU	TER
GIZA++	19.48	30.13	56.63
BITG	20.35	34.03	50.94
+ SRL_{MEANT}	**20.43**	**34.27**	**50.35**

properly since it is based on generalization. We use MEANT to confidence-weight the training data for the Tigrinya–English translation task then shuffle the data so that the identical sentence pairs are not in the same batch. Table 5 shows that the SMT model highly outperforms the NMT model for both the unbiased models and the MEANT constrained models. The results might seem very low for an NMT model, but, we highlight the point that to maintain the apples-to-apples low-resource generalization comparison we are using raw data without any preprocessing and without any additional high-resource dependent techniques like knowledge transfer from similar high-resource languages.

6.3 Our models also perform well in a high resource setting

We tested the MEANT based model with Latvian–English translation task (results in table 6), which is not low resource in this case since it has more than 600K sentence pairs. Table 6 shows that our approach slightly improves the translation quality compared to BITGs, but highly outperforms GIZA++ based model. This shows that, although our novel approach improves the MT quality in a high resource setup, it definitely has a higher impact when dealing with low resource languages.

6.4 Translation examples

In example 1 (figure 3), the MEANT based model produces a translation that is as good as the reference. However, both BITG and GIZA++ based translations completely fail to capture the word opera. Example 2 (figure 3) is from the Turkish–English translation task. In this example, the MEANT based model only fails at translating the name of the city Belede; otherwise, the translation sounds better than the two other systems. The BITG model output has Yangon, which does not appear in the Turkish input (see gloss).

7 Conclusion

We have shown that adopting the monolingual semantic evaluation metric MEANT as an objective function for driving ITG induction yields a high improvement compared to the conventional alignment methods on many challenging low resource languages. We have also proposed another heuristic for evaluating how good an English semantic parse is, then used it to induce ITGs. We have experimented with several challenging low resource languages from different language families and have demonstrated that using a monolingual semantic frame based objective function during the actual learning of the translation model helps learn good bilingual correlations with a relatively small dataset in contrast to conventional SMT systems. The promising results we report in this new line of research make it seem that learning more semantically motivated translation models might be less challenging than generally assumed and is worth exploring.

8 Acknowledgment

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

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.

Wilker Aziz, Miguel Rios, and Lucia Specia. Shallow semantic trees for SMT. In *Sixth Workshop on Statistical Machine Translation (WMT 2011)*, 2011.

Meriem Beloucif and Dekai Wu. Driving inversion transduction grammar induction with semantic evaluation. In *5th Joint Conference on Lexical and Computational Semantics at ACL*, 2016.

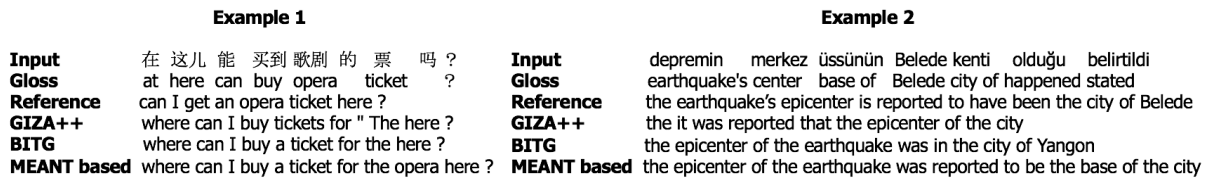

Figure 3: Examples comparing the output from the three discussed alignment systems extracted from the Chinese–English and the Turkish–English translation tasks.

Meriem Beloucif and Dekai Wu. Improving word alignment for low resource languages using english monolingual srl. In *Sixth Workshop on Hybrid Approaches to Translation (HyTra-6) at COLING. Osaka, Japan*, 2016.

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.

Anders Björkelund, Love Hafdell, and Pierre Nugues. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, Colorado, June 2009. Association for Computational Linguistics.

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.

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

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

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.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *The Tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand, September 2005.

Mamoru Komachi, Yuji Matsumoto, and Masaaki Nagata. Phrase reordering for statistical machine translation based on predicate-argument structure. In *International Workshop on Spoken Language Translation (IWSLT 2006)*, 2006.

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. Semantic role features for machine translation. In *23rd International Conference on Computational Linguistics (COLING 2010)*, 2010.

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

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.

Minh-Thang Luong and Christopher D Manning. Stanford neural machine translation systems for spoken language domains. In *The International Workshop on Spoken Language Translation (IWSLT15)*, 2015.

Minh-Thang Luong and Christopher D Manning. Achieving open vocabulary neural machine translationwith hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 1054–1063, 2016.

Graham Neubig. lamtram: A toolkit for language and translation modeling using neural networks, 2015.

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.

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.

Michael Roth and Kristian Woodsend. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 407–413,

Doha, Qatar, October 2014. Association for Computational Linguistics.

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 and Dekai Wu. Reestimation of reified rules in semiring parsing and biparsing. In *Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-5)*, pages 70–78, Portland, Oregon, June 2011. Association for Computational Linguistics.

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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. volume abs/1508.07909, 2015.

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.

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.

Christoph Tillmann, Stephan Vogel, Hermann Ney, Arkaitz Zubiaga, and Hassan Sawaf. Accelerated DP based search for statistical translation. In *Fifth European Conference on Speech Communication and Technology (EUROSPEECH 1997)*, 1997.

Dekai Wu and Pascale Fung. Semantic roles for SMT: A hybrid two-pass model. In *Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2009)*, pages 13–16, 2009.

Xianchao Wu, Katsuhito Sudoh, Kevin Duh, Hajime Tsukada, and Masaaki Nagata. Extracting preordering rules from predicate-argument structures. In *The 5th International Joint Conference on Natural Language Processing (IJCNLP2011)*, 2011.

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

Deyi Xiong, Min Zhang, and Haizhou Li. Modeling the translation of predicate-argument structure for SMT. In *50th Annual Meeting of the Association for Computational Linguistics (ACL 2012)*, 2012.