# Driving inversion transduction grammar induction with semantic evaluation

**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

We describe a new technique for improving statistical machine translation training by adopting scores from a recent crosslingual semantic frame based evaluation metric, XMEANT, as outside probabilities in expectation-maximization based ITG (inversion transduction grammars) alignment. Our new approach strongly biases early-stage SMT learning towards semantically valid alignments. Unlike previous attempts that have proposed using semantic frame based evaluation metrics as the objective function for late-stage tuning of less than a dozen loglinear mixture weights, our approach instead applies the semantic metric at one of the earliest stages of SMT training, where it may impact millions of model parameters. The choice of XMEANT is motivated by empirical studies that have shown ITG constraints to cover almost all crosslingual semantic frame alternations, which resemble the crosslingual semantic frame matching measured by XMEANT. Our experiments purposely restrict training data to small amounts to show the technique's utility in the absence of a huge corpus, to study the effects of semantic generalizations while avoiding overreliance on memorization. Results show that directly driving ITG training with the crosslingual semantic frame based objective function not only helps to further sharpen the ITG constraints, but still avoids excising relevant portions of the search space, and leads to better performance than either conventional ITG or GIZA++ based approaches.

## 1 Introduction

We propose a new technique that biases early stage statistical machine translation (SMT) learning towards semantics. Our algorithm adopts the crosslingual evaluation metric XMEANT (Lo *et al.*, 2014) to initialize expectation-maximization (EM) outside probabilities during inversion transduction grammar or ITG (Wu, 1997) induction. We show that injecting a crosslingual semantic frame based objective function in the actual learning of the translation model helps to bias the training of the SMT model towards semantically more relevant structures. Our approach is highly motivated by recent research which showed that including a semantic frame based objective function during the formal feature weights tuning stage increases the translation quality. More precisely, Lo *et al.* (2013a); Lo and Wu (2013); Lo *et al.* (2013b); Beloucif *et al.* (2014) showed that tuning against a semantic frame based evaluation metric like MEANT (Lo *et al.*, 2012), improves the translation adequacy.

Our choice to improve ITG alignments is motivated by the fact that they have already previously been empirically shown to cover essentially 100% of crosslingual semantic frame alternations, even though they rule out the majority of incorrect alignments (Addanki *et al.*, 2012). Our technique uses XMEANT for rewarding good translations while learning bilingual correlations of the translation model. We also show that integrating a semantic frame based objective function much earlier in the training pipeline not only produces more semantically correct alignments but also helps to learn bilingual correlations without memorizing from a huge amounts of parallel corpora. We report results and examples showing that this way for inducing ITGs gives a better translation quality compared to the conventional ITGs and GIZA++ (Och

*Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (*SEM 2016)*, pages 55–63,
Berlin, Germany, August 11-12, 2016.

and Ney, 2000) alignments.

## 2 Related work

The choice of XMEANT, a crosslingual version of MEANT (Lo and Wu, 2011, 2012; Lo *et al.*, 2012), is motivated by the work of Lo *et al.* (2014) who showed that XMEANT can correlate better with human adequacy judgement than most other metrics under some conditions. Furthermore, previous empirical studies have shown that the crosslingual semantic frame matching measured by XMEANT is fully covered within ITG constraints (Addanki *et al.*, 2012).

### 2.1 Inversion transduction grammars

Inversion transduction grammars (ITGs, Wu (1997)) are a subset of syntax-directed transduction grammar (Lewis and Stearns, 1968; Aho and Ullman, 1972). A transduction is a set of bisentences that define the relation between an input language $L_0$ and an output language $L_1$. Accordingly, a transduction grammar is able to generate, translate or accept a transduction or a set of bisentences. Inversion transductions are a subset of transduction which are synchronously generated and parsed by inversion transduction grammars (ITGs, (Wu, 1997)).

An ITG can always be written in a 2-normal form and it is represented by a tuple $\langle N, V_0, V_1, R, S\rangle$ where $N$ is a set of nonterminals, $V_0$ and $V_1$ are the bitokens of $L_0$ and $L_1$ respectively, $R$ is a set of transduction rules and $S \in N$ is the start symbol.

We can write each transduction rule as follows:

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

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

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

There are different classes of inversion transduction grammars. LTGs or linear transduction grammars (Saers *et al.*, 2010) impose harsher constraints than ITGs but still cover almost 100% of verb frame alternations (Addanki *et al.*, 2012). There are also many ways to formulate the model over ITGs: Wu (1995); Zhang and Gildea (2005); Chiang (2007); Cherry and Lin (2007); Blunsom *et al.* (2009); Haghighi *et al.* (2009); Saers *et al.* (2010); Neubig *et al.* (2011).

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

### 2.2 Semantic frame based evaluation metrics

#### 2.2.1 MEANT's algorithm

Unlike *n*-gram or edit-distance based metrics, the MEANT family of metrics (Lo and Wu, 2011, 2012; Lo *et al.*, 2012) adopt the principle that a good translation is one in which humans can successfully understand the general meaning of the input sentence as captured by the basic event structure: *who did what to whom, for whom, when, where, how and why* (Pradhan *et al.*, 2004). Recent work have shown that the semantic frame based metric, MEANT, correlates better with human adequacy judgment than most common evaluation metrics (Lo and Wu, 2011, 2012; Lo *et al.*, 2012) such as BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006).

Algorithm one in figure 2 shows how a MEANT score is computed (Lo and Wu, 2011, 2012; Lo *et al.*, 2012).

#### 2.2.2 XMEANT: crosslingual MEANT

XMEANT (Lo *et al.*, 2014) is the crosslingual version of the semantic evaluation metric MEANT. It has been shown that the crosslingual evaluation metric, XMEANT, correlates even better with human adequacy judgment than MEANT, and also better than most evaluation metrics like BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002), METEOR (Banerjee and Lavie, 2005), CDER (Leusch *et al.*, 2006), WER (Nießen *et al.*, 2000), and TER (Snover *et al.*, 2006).

Unlike MEANT which needs expensive man-made references, XMEANT uses the foreign in-

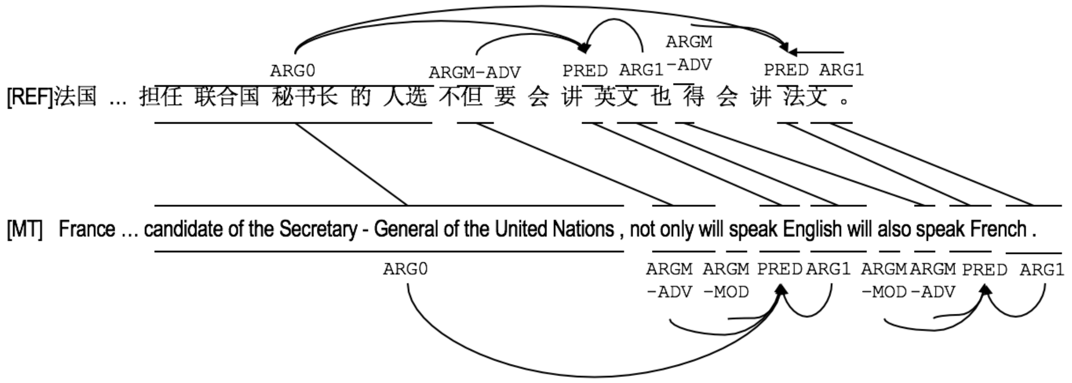

ARG0
ARGM-ADV
PRED
ARG1
ARGM -ADV
PRED
ARG1
[REF]法国 ... 担任 联合国 秘书长 的 人选 不但 要 会 讲 英文 也 得 会 讲 法文 。
[MT] France ... candidate of the Secretary - General of the United Nations , not only will speak English will also speak French .
ARG0
ARGM -ADV
ARGM -MOD
PRED
ARG1
ARGM -MOD
ARGM -ADV
PRED
ARG1

Figure 1: Example of how XMEANT aligns words and phrases

put to evaluate the MT translation output. Figure 1 shows an example of shallow semantic parsing in a Chinese input sentence and an English MT output. It also shows how XMEANT aligns the role fillers between two parallel sentences from different languages based on their semantic frames matching.

Figure 2 underlines the differences between MEANT and XMEANT algorithms. XMEANT uses MEANT's f-score based method for aggregating lexical translation probabilities within semantic role filler phrases. Each token of the role fillers in the output/input string is aligned to the token of the role fillers in the input/output string that has the maximum lexical translation probability. In contrast to MEANT which measures lexical similarity using a monolingual context vector model, XMEANT instead substitutes simple crosslingual lexical translation probabilities. The crosslingual phrasal similarities are computed as follows:

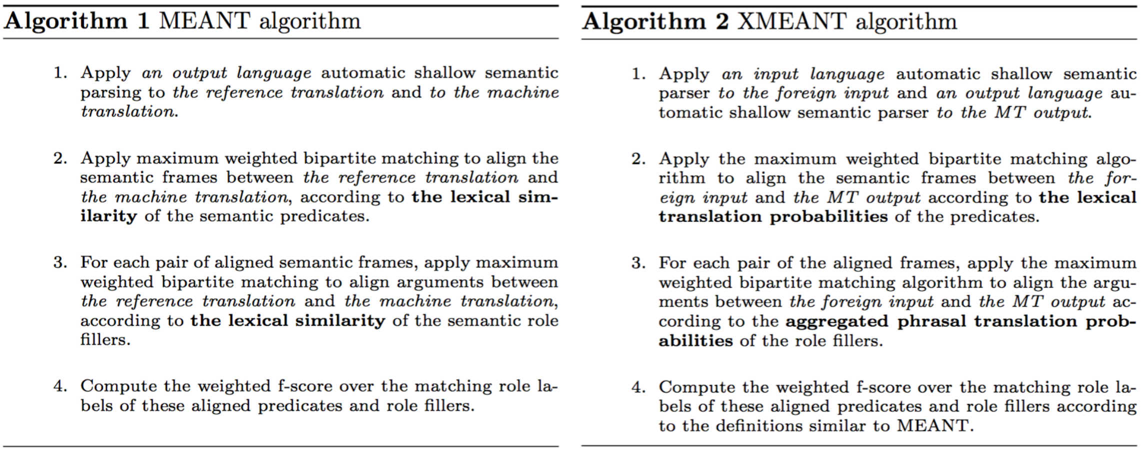
$$
\begin{aligned}
\mathbf{e}_{i,\text{pred}} &\equiv \text{the output side of the pred of aligned frame } i \\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the input side of the pred of aligned frame } i \\
\mathbf{e}_{i,j} &\equiv \text{the output side of the ARG } j \text{ of aligned frame } i \\
\mathbf{f}_{i,j}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

where the joint probability $p$ is defined as the harmonic mean of the two directions of the translation table $t$ trained using IBM model 1 (Brown *et al.*, 1993). $\text{prec}_{\mathbf{e},\mathbf{f}}$ is the precision and $\text{rec}_{\mathbf{e},\mathbf{f}}$ is the recall of the phrasal similarities of the role fillers. $s_{i,\text{pred}}$ and $s_{i,j}$ are the f-scores of the phrasal similarities of the predicates and role fillers of the arguments of type $j$ between the input and the MT output.

Our approach uses the XMEANT score of every bisentence in the training data and uses it to initialize the outside probability of the expectation-maximization algorithm, then uses this crucial information for weighting meaningful sentences to inducing bracketing inversion transduction grammars. We show in this paper that using this semantic objective function at an early stage of training SMT system, we are not only able to learn more semantic bilingual correlations between the two languages, but we are also able get rid of the heavy memorization that most of the conventional alignment systems rely heavily on.

### 2.3 Alignment

Word alignment is considered to be a necessary step in training SMT systems, it helps to learn bilingual correlations between the input and the output languages. In this work, we compare the alignment produced by our system to the traditional GIZA++ alignment and the conventional ITG alignment. Most of the conventional alignment algorithms: IBM models (Brown *et al.*, 1990) and hidden Markov models or HMM (Vogel *et al.*, 1996) are flat and directed. In fact, (a) they allow the unstructured movement of words leading to a weak word alignment, (b) consider translations in one direction in isolation, and (c)



**Algorithm 1** MEANT algorithm

1. Apply *an output language* automatic shallow semantic parsing to *the reference translation* and *to the machine translation.*
2. Apply maximum weighted bipartite matching to align the semantic frames between *the reference translation* and *the machine translation*, according to **the lexical similarity** of the semantic predicates.
3. For each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between *the reference translation* and *the machine translation*, according to **the lexical similarity** of the semantic role fillers.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers.

**Algorithm 2** XMEANT algorithm

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

Figure 2: MEANT vs XMEANT algorithms

need two separate alignments to form a single *bidirectional alignment*. The harmonization of two directed alignments is typically done heuristically. This means that there is no model that considers the final bidirectional alignment that the translation system is trained on to be optimal. Inversion transduction grammars (Wu, 1997), on the other hand, have proven that learning word alignments using a system that is compositionally-structured, can provide *optimal bidirectional alignments*. Although this structured optimality comes at a higher cost in terms of time complexity, it allows preexisting structured information to be incorporated into the model. It also allows models to be compared in a meaningful way. Saers and Wu (2009) proposed a better method of producing word alignment by training inversion transduction grammars (Wu, 1997). One problem encountered with such model was that the exhaustive biparsing that runs in $O(n^6)$. Saers *et al.* (2009) proposed a more efficient algorithm that runs in $O(n^3)$.

Zens and Ney (2003) showed that ITG constraints allow a higher flexibility in word-ordering for longer sentences than the conventional IBM model. Furthermore, they demonstrate that applying ITG constraints for word alignment leads to learning a significantly better alignment than the constraints used in conventional IBM models for both German-English and French-English language pairs. Zhang and Gildea (2005) on the other hand showed that the tree learned while training using ITG constraints gives much more accurate word alignments than those trained on manually annotated treebanks like in Yamada and Knight (2001) in both Chinese-English and German-English. Haghighi *et al.* (2009) show that using ITG constraints for supervised word alignment methods not only produce alignments without lower alignment error rates but also produces a better translation quality.

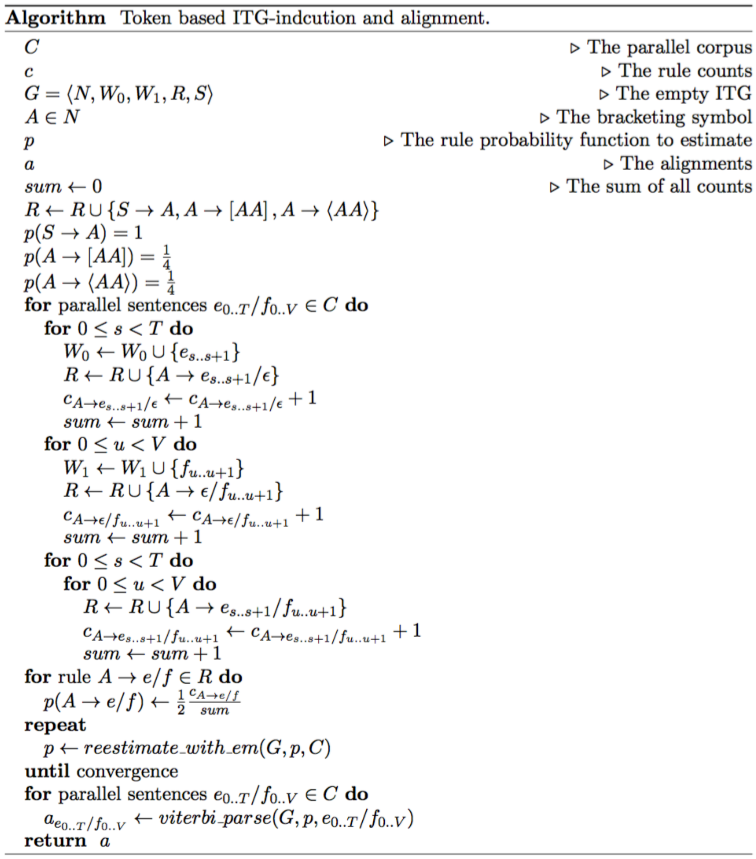
Some of the previous work on word alignment used morphological and syntactic features (De Gispert *et al.*, 2006). Log linear models have been proposed to incorporate those features (Chris *et al.*, 2011). The problem with those approaches is that they require language specific knowledge and they always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

A few studies that try to integrate some semantic knowledge in computing word alignment are proposed by Jeff *et al.* (2011) and Theerawat and David (2014). However, the former needs to have a prior word alignment learned on lexical items. The latter proposes a semantically oriented word alignment, but requires extracting word similarities from the monolingual data first, before producing alignment using word similarities.

## 3 Adopting XMEANT scores as EM outside probabilities

We implemented a token based BITG system as our ITG baseline, our choice of BITG is motivated by previous work that showed that BITG alignments outperformed alignments from GIZA++ (Saers *et al.*, 2009).

Figure 3 shows the BITG induction algorithm that we used in this paper. We initialize it with



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

Figure 3: Token based BITG induction algorithm

Table 1: Comparison of translation quality for three methods used to train Moses for Chinese-English MT under small corpus IWSLT 2007 conditions

| | cased | | uncased | |
|---|---|---|---|---|
| System | BLEU | TER | BLEU | TER |
| Giza++ based induction | 19.23 | 63.94 | 19.83 | 63.40 |
| ITG based induction | 20.05 | 63.19 | 20.42 | 62.61 |
| XMEANT outside probabilities based | **27.59** | **59.48** | **28.54** | **58.81** |

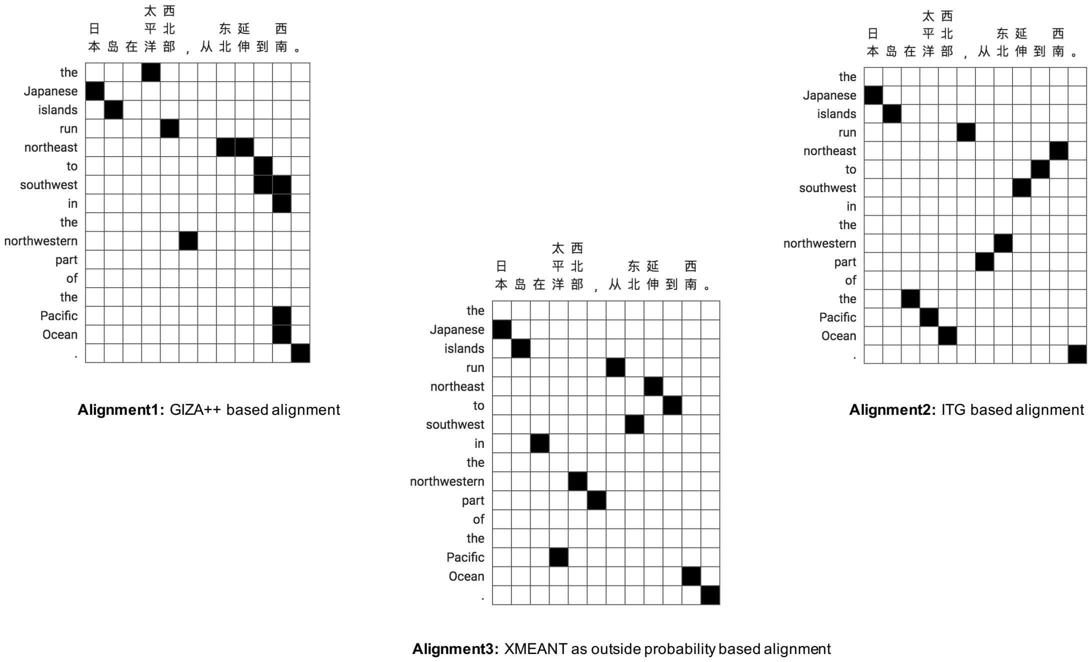
uniform structural probabilities, setting aside half of the probability mass for lexical rules. This probability mass is distributed among the lexical rules according to co-occurrence counts from the training data, assuming each sentence to contain one empty token to account for singletons. The novelty in our model consists of adopting the XMEANT score of each bisentence as the initial value for the outside probabilities as follows:

$$\beta_{(0,|\mathbf{e}_i|,0,|\mathbf{f}_i|)} = XMEANT(\mathbf{e}_i, \mathbf{f}_i) \qquad (1)$$

where $i$ represents the bisentences number $i$ in the corpus.

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.

In our experiments, we tried to show that including semantic earlier in learning SMT systems can help us get rid of the expensive huge corpora used in the traditional SMT training. Although Chinese is not a low resource language, we tried purposely to simulate low resource conditions, we used a relatively small corpus (IWSLT07). The training set contained 39,953 sentences. The dev set and test set were the same for all systems in order to keep


日本 岛 在 太平洋 西北部 ， 从 东北 延伸 到 西南 。
the Japanese islands run northeast to southwest in the northwestern part of the Pacific Ocean .

**Alignment1:** GIZA++ based alignment

**Alignment3:** XMEANT as outside probability based alignment

**Alignment2:** ITG based alignment

Figure 4: Alignments of bisentences produced by the three discussed alignment systems

the experiments comparable.

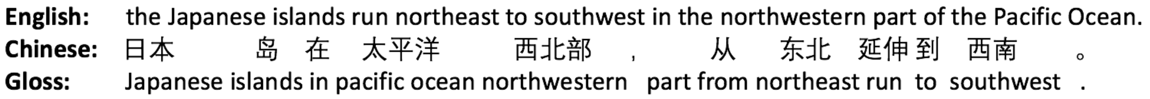
We compare the performance of our proposed semantic frame based alignment to the conventional ITG alignment and to the traditional GIZA++ baseline with grow-diag-final-and to harmonize both alignment directions. We tested the different alignments described above by using the standard Moses toolkit (Koehn *et al.*, 2007), and a 6-gram language model learned with the SRI language model toolkit (Stolcke, 2002) to train our model.

## 4 Results

We compared the performance of the semantic frame based ITG alignment against both the conventional ITG alignment and the traditional GIZA++ alignment. We evaluated our MT output using the surface based evaluation metric BLEU (Papineni *et al.*, 2002) and the edit distance evaluation metric TER (Snover *et al.*, 2006). Table 1 shows that the alignment based on our proposed algorithm helps achieving much higher scores in term of BLEU and TER in comparison to both conventional ITG and GIZA++ alignment.

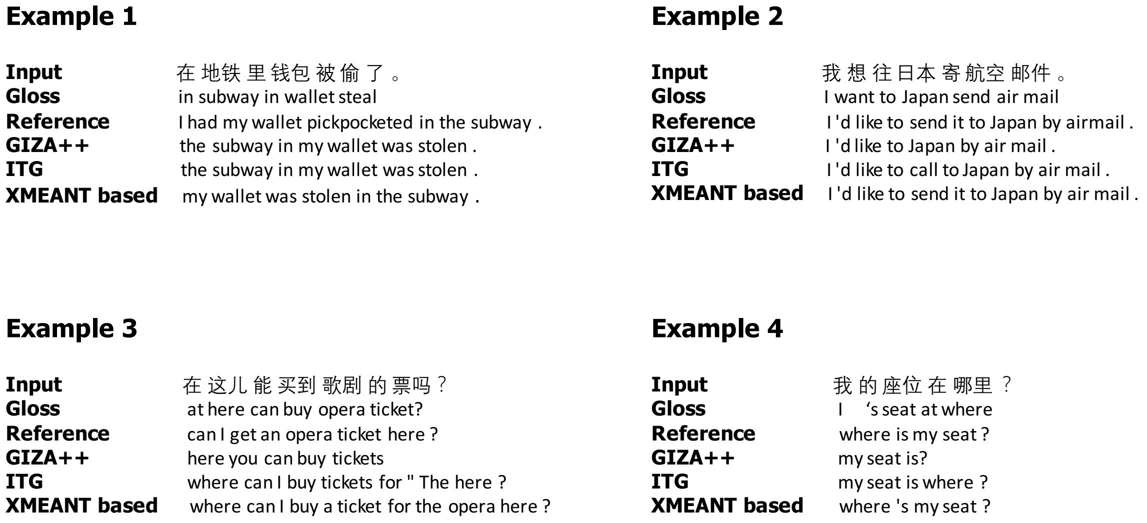
Figure 4 illustrates the alignments generated by the three systems described in this paper for a given example. The traditional GIZA++ alignment (top left) and the conventional ITG alignment (top right) fail to align all the crucial parts of the given bisentence. The English sentence can be divided into three major parts: “the Japanese islands”, “run northeast to southwest” and “in the northwest part of the pacific ocean.”. The conventional ITG based alignment only succeeds to align the first part of the sentence. GIZA++ based system correctly aligns part one and parts of part two. We note from the sentence’s gloss (figure 5) that our proposed alignment outperforms the two other alignments by capturing the relevant information in both part one and part three, and also successfully aligns the token “in” to “在”.

Figure 6 shows four interesting examples extracted from our translated data and compared to the translations obtained by other systems. We see from the examples that ITG based models can produce a slightly better outputs compared to GIZA++ based alignment, but our semantic frame based alignment highly outperform both alignments. We clearly see how the outputs from our new submitted system capture more strong bilingual correlations although we are using the same small corpus for every system. In example 2 and 4, our system produces a translation that is as good as the human reference. For example number one, our system produces a more precise translation than the human reference since the Chinese character “偷” is normally translated to “stolen” and not “pickpocketed”. Example 3, our proposed system



| | |
|---|---|
| **English:** | the Japanese islands run northeast to southwest in the northwestern part of the Pacific Ocean. |
| **Chinese:** | 日本 岛 在 太平洋 西北部 ， 从 东北 延伸 到 西南 。 |
| **Gloss:** | Japanese islands in pacific ocean northwestern part from northeast run to southwest . |

Figure 5: The gloss of the bisentence used in figure 4



**Example 1**

| | |
|---|---|
| **Input** | 在 地铁 里 钱包 被 偷 了 。 |
| **Gloss** | in subway in wallet steal |
| **Reference** | I had my wallet pickpocketed in the subway . |
| **GIZA++** | the subway in my wallet was stolen . |
| **ITG** | the subway in my wallet was stolen . |
| **XMEANT based** | my wallet was stolen in the subway . |

**Example 2**

| | |
|---|---|
| **Input** | 我 想 往 日本 寄 航空 邮件 。 |
| **Gloss** | I want to Japan send air mail |
| **Reference** | I 'd like to send it to Japan by airmail . |
| **GIZA++** | I 'd like to Japan by air mail . |
| **ITG** | I 'd like to call to Japan by air mail . |
| **XMEANT based** | I 'd like to send it to Japan by air mail . |

**Example 3**

| | |
|---|---|
| **Input** | 在 这儿 能 买到 歌剧 的 票吗 ? |
| **Gloss** | at here can buy opera ticket? |
| **Reference** | can I get an opera ticket here ? |
| **GIZA++** | here you can buy tickets |
| **ITG** | where can I buy tickets for " The here ? |
| **XMEANT based** | where can I buy a ticket for the opera here ? |

**Example 4**

| | |
|---|---|
| **Input** | 我 的 座位 在 哪里 ? |
| **Gloss** | I 's seat at where |
| **Reference** | where is my seat ? |
| **GIZA++** | my seat is? |
| **ITG** | my seat is where ? |
| **XMEANT based** | where 's my seat ? |

Figure 6: Four interesting examples comparing the output from the three discussed alignment systems

give the most accurate and understandable translation among all systems. The only small problem with this output is the fact that the Chinese character "在" which represents "at" but sometimes gets translated to "where".

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

## 5 Conclusion

We presented a semantic frame based alignment method that adopts the crosslingual semantic evaluation metric, XMEANT, as expectation maximization (EM) outside probabilities for inversion transduction grammar (ITG) induction. We show that our new approach biases early stage SMT training towards semantics by injecting a semantic frame objective function in the initial steps of learning the translation model. Incorporating the semantic frame based objective function at the early stage of induction biases ITG alignments at a point where it still has the potential to influence millions of model parameters. Finally, we show that directly driving ITG induction with a crosslingual semantic frame objective function not only helps to further sharpen the ITG constraints, but still avoids excising relevant portions of the search space, and leads to better performance than either conventional ITG or GIZA++ based approaches.

## 6 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.

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

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

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

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

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederik Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85, 1990.

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

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

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

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

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

George Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *The second international conference on Human Language Technology Research (HLT '02)*, San Diego, California, 2002.

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

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

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

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

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

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

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

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

Chi-kiu Lo, 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.

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

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. A evaluation tool for machine translation: Fast evaluation for MT research. In *The Second International Con-*

*ference 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.

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

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

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

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

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.

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

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

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

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

Kenji Yamada and Kevin Knight. A syntax-based statistical translation model. In *39th Annual Meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 523–530, Toulouse, France, July 2001.

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

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