IMEANT
Better semantic frame based MT evaluation via inversion transduction grammars

Dekai WU   LO Chi-kiu  Meriem BELOUCIF  Markus SAERS
{ dekai | jackielo | mbeloucif | masaers }@cs.ust.hk

HKUST
Human Language Technology Center
Department of Computer Science and Engineering
University of Science and Technology, Hong Kong
what makes a translation useful

how well is

who did what to whom, for whom, when, where, why and how

preserved in translation?
how well do n-grams match between reference and machine translations?
how well do semantic frames match between reference and machine translations?
HMEANT
Human semantic MT evaluation via SRL

[IN] 至此，在中国内地停售近两个月的SK-II全线产品恢复销售。

[REF] Until after their sales had ceased in mainland China for almost two months, sales of the complete range of SK-II products have now be resumed.

[MT1] So far, nearly two months SK-II the sale of products in the mainland of China to resume sales.

[MT2] So far, in the mainland of China to stop selling nearly two months of SK-2 products sales resumed.

[MT3] So far, the sale in the mainland of China for nearly two months of SK-II line of products.
Example: a less useful translation
Fewer SRL matches 😊
but more N-gram and syntax-subtree matches! 😞

So far, the sale in the mainland of China for nearly two months of SK-II line of products.
Conversely: a more useful translation
More SRL matches 😊
but fewer N-gram and syntax-subtree matches! 😞
HMEANT is just an f-score on semantic frame match (with a tiny number of weights)

- **sentence accuracy**: avg translation accuracy over all frames of a sentence
  sentence precision (or recall) = frame precision (or recall) averaged across the total number of frames in MT (or REF)

- **frame accuracy**: avg translation accuracy over all roles of a frame
  frame precision (or recall) = weighted sum of # correctly translated arguments, normalized by the weighted sum of # arguments in MT (or REF)

- **frame importance**: weight each frame by its span coverage ratio

- **role importance**: weight each type of role by maximizing HMEANT’s correlation with HAJ using a human ranked training corpus

\[
\begin{align*}
  m_i & \equiv \frac{\text{# word tokens filled in Frame } i}{\text{total # word tokens in MT}} \\
  r_i & \equiv \frac{\text{# word tokens filled in Frame } i}{\text{total # word tokens in REF}} \\
  \text{precision} & = \frac{\sum_i m_i \cdot w_{\text{pred}} + \sum_{j} w_j (C_{i,j} + w_{\text{partial}} R_{i,j})}{\sum_i m_i} \\
  \text{recall} & = \frac{\sum_i r_i \cdot w_{\text{pred}} + \sum_{j} w_j (C_{i,j} + w_{\text{partial}} R_{i,j})}{\sum_i r_i}
\end{align*}
\]
**HMEANT, MEANT, UMEANT**  a family of semantic frame based MT evaluation metrics

- **HMEANT** human  [Lo & Wu, ACL, IJCAI, SSST 2011]
  - assesses MT utility via semantic frames with high representational transparency
  - needs only unskilled humans to annotate and align semantic frames
  - correlates with human adequacy judgment better than HTER at lower labor cost
  - applies easily on any language pair

- **MEANT** automatic  [Lo, Tumuluru & Wu, WMT 2012]
  - outperforms all commonly used automatic MT evaluation metrics
    - replaces human SRL with automatic shallow semantic parsing
    - replaces human semantic frame alignment with automatic alignment
  - simple & transparent – preserves Occam’s razor spirit of HMEANT
  - now in both English and Chinese
  - top 4 in WMT2013 metrics track evaluation

- **UMEANT** unsupervised automatic  [Lo & Wu, SSST 2012]
  - eliminates any dependency on a corpus with human ranked MT output in training the weights of semantic role labels by estimating them via the relative frequency of the labels in the reference
  - good for resource-sparse languages
  - top 3 in WMT2013 metrics track evaluation
the first ever
directly semantically trained SMT systems

- **why tune MT against MEANT?**
  - produces more robustly adequate translations than tuning against BLEU or TER
    - across genres (newswire, web forum, TED)
    - across output languages (English, Chinese)
    - across MT paradigms (phrase based, hierarchical phrase based)
  - constrains the MT system to make more accurate lexical and reordering choices
    - preserving the meaning of the translation as captured by semantic frames right in the training process
  - the first time in 25 years of history that SMT has ever been directly trained to **maximize preserving who did what to whom, for whom, when, where, how, why** (a bit scary!)
XMEANT a cross-lingual semantic frame based MT evaluation metric

- **XMEANT** cross-lingual MEANT [Lo, Beloucif, Saers & Wu, ACL 2014]

  - eliminates the need for expensive reference translations ...
    yet correlates with human adequacy judgment even more closely than MEANT!

  - since words come from different vocabularies for input and output languages, can’t use MEANT’s word vector similarities to align role fillers any more; instead use translation probabilities plus **language-independent BITGs constraints** (Wu 1997; Zens & Ney 2003; Saers & Wu 2009)

  - a new generation of Wu & Fung’s (NAACL, EAMT 2009) cross-lingual score ... that exploits all our recent advances on monolingual MEANT

- well, if BITG constraints work so well for cross-lingual XMEANT... could they also improve ordinary monolingual MEANT?
IMEANT  

new! an ITG-based semantic frame based MT evaluation metric

- **further improves** MEANT’s correlation with human adequacy judgment which was already high

- achieved by using **bracketing ITGs** to biparse the semantic role fillers in both reference and machine translations

- shows that ITGs
  - appropriately constrain the allowable permutations between the compositional segments across the reference and machine translations
  - score the phrasal similarity of the semantic role fillers more accurately than the simple heuristics like bag-of-word alignment or maximum alignment
1. apply automatic shallow semantic parsing to the reference and machine translations

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
1. apply automatic shallow semantic parsing to the reference and machine translations

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

\[
\begin{align*}
q_{i,j}^0 & \equiv \text{ARG } j \text{ of aligned frame } i \text{ in MT} \\
q_{i,j}^1 & \equiv \text{ARG } j \text{ of aligned frame } i \text{ in REF} \\
w_i^0 & \equiv \frac{\text{#tokens filled in aligned frame } i \text{ of MT}}{\text{total #tokens in MT}} \\
w_i^1 & \equiv \frac{\text{#tokens filled in aligned frame } i \text{ of REF}}{\text{total #tokens in REF}} \\
w_{\text{pred}} & \equiv \text{weight of similarity of predicates} \\
w_j & \equiv \text{weight of similarity of ARG } j \\
s_{i,\text{pred}} & \equiv \text{predicate similarity in aligned frame } i \\
s_{i,j} & \equiv \text{ARG } j \text{ similarity in aligned frame } i \\
\text{precision} & = \frac{\sum_i w_i^0 \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q_{i,j}^0|}}{\sum_i w_i^0} \\
\text{recall} & = \frac{\sum_i w_i^1 \frac{w_{\text{pred}} s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q_{i,j}^1|}}{\sum_i w_i^1} \\
\text{MEANT} & = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}}
\end{align*}
\]
1. apply automatic shallow semantic parsing to the reference and machine translations

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
MEANT

\[ q_i^0 = \text{ARG } j \text{ of aligned frame } i \text{ in MT} \]
\[ q_i^1 = \text{ARG } j \text{ of aligned frame } i \text{ in REF} \]
\[ w_i^0 = \frac{\text{# tokens filled in aligned frame } i \text{ of MT}}{\text{total # tokens in MT}} \]
\[ w_i^1 = \frac{\text{# tokens filled in aligned frame } i \text{ of REF}}{\text{total # tokens in REF}} \]
\[ w_{\text{pred}} = \text{weight of similarity of predicates} \]
\[ w_j = \text{weight of similarity of } \text{ARG } j \]
\[ e_{i,\text{pred}} = \text{pred string of the aligned frame } i \text{ of MT} \]
\[ f_{i,\text{pred}} = \text{pred string of the aligned frame } i \text{ of REF} \]
\[ e_{i,j} = \text{role fillers of } \text{ARG } j \text{ of the aligned frame } i \text{ of MT} \]
\[ f_{i,j} = \text{role fillers of } \text{ARG } j \text{ of the aligned frame } i \text{ of REF} \]
\[ s(e, f) = \text{lexical similarity of token } e \text{ and } f \]

\[ \begin{align*}
\text{prec}_{e,f} &= \frac{\Sigma_{e \in e} \max s(e,f)}{|e|} \\
\text{rec}_{e,f} &= \frac{\Sigma_{f \in f} \max s(e,f)}{|f|} \\
\text{s}_{i,\text{pred}} &= \frac{2 \times \text{prec}_{e_{i,\text{pred}},f_{i,\text{pred}}} \times \text{rec}_{e_{i,\text{pred}},f_{i,\text{pred}}}}{\text{prec}_{e_{i,\text{pred}},f_{i,\text{pred}}} + \text{rec}_{e_{i,\text{pred}},f_{i,\text{pred}}}} \\
\text{s}_{i,j} &= \frac{2 \times \text{prec}_{e_{i,j},f_{i,j}} \times \text{rec}_{e_{i,j},f_{i,j}}}{\text{prec}_{e_{i,j},f_{i,j}} + \text{rec}_{e_{i,j},f_{i,j}}} \\
\text{precision} &= \frac{\Sigma_i w_i^0 \text{w}_{\text{pred}} s_{i,\text{pred}}^0 + \Sigma_j w_j^0 s_{j,i}^0}{\Sigma_i w_i^0} \\
\text{recall} &= \frac{\Sigma_i w_i^1 \text{w}_{\text{pred}} s_{i,\text{pred}}^1 + \Sigma_j w_j^1 s_{j,i}^1}{\Sigma_i w_i^1} \\
\text{MEANT} &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\end{align*} \]

IMEANT

\[ q_i^0 = \text{ARG } j \text{ of aligned frame } i \text{ in MT} \]
\[ q_i^1 = \text{ARG } j \text{ of aligned frame } i \text{ in REF} \]
\[ w_i^0 = \frac{\text{# tokens filled in aligned frame } i \text{ of MT}}{\text{total # tokens in MT}} \]
\[ w_i^1 = \frac{\text{# tokens filled in aligned frame } i \text{ of REF}}{\text{total # tokens in REF}} \]
\[ w_{\text{pred}} = \text{weight of similarity of predicates} \]
\[ w_j = \text{weight of similarity of } \text{ARG } j \]
\[ e_{i,\text{pred}} = \text{pred string of the aligned frame } i \text{ of MT} \]
\[ f_{i,\text{pred}} = \text{pred string of the aligned frame } i \text{ of REF} \]
\[ e_{i,j} = \text{role fillers of } \text{ARG } j \text{ of the aligned frame } i \text{ of MT} \]
\[ f_{i,j} = \text{role fillers of } \text{ARG } j \text{ of the aligned frame } i \text{ of REF} \]
\[ s(e, f) = \text{lexical similarity of token } e \text{ and } f \]
\[ G \equiv \{ (A), \emptyset^0, \emptyset^1, \emptyset, A \} \]
\[ \emptyset \equiv \{ A \rightarrow [A], A \rightarrow \langle A, A \rangle, A \rightarrow e/f \} \]
\[ p((A A)|A) = p((A A)|A) = 1 \]
\[ p(e/f | A) = s(e,f) \]
\[ s_{i,\text{pred}} = \lg^{-1} \left( \frac{\lg \left( P(A \rightarrow e_{i,\text{pred}}/f_{i,\text{pred}} | G) \right)}{\max(\{|e_{i,\text{pred}}|, |f_{i,\text{pred}}|\})} \right) \]
\[ s_{i,j} = \lg^{-1} \left( \frac{\lg \left( P(A \rightarrow e_{i,j}/f_{i,j} | G) \right)}{\max(\{|e_{i,j}|, |f_{i,j}|\})} \right) \]

\[ \begin{align*}
\text{precision} &= \frac{\Sigma_i w_i^0 \text{w}_{\text{pred}} s_{i,\text{pred}}^0 + \Sigma_j w_j^0 s_{j,i}^0}{\Sigma_i w_i^0} \\
\text{recall} &= \frac{\Sigma_i w_i^1 \text{w}_{\text{pred}} s_{i,\text{pred}}^1 + \Sigma_j w_j^1 s_{j,i}^1}{\Sigma_i w_i^1} \\
\text{IMEANT} &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\end{align*} \]
IMEANT outperforms the most recent version of MEANT

- IMEANT shows a 3 point improvement over MEANT on GALE-A
- IMEANT is tied with MEANT in correlation with HAJ on GALE-B

Table 1. Sent-level correlation with HAJ on GALE P2.5 data

<table>
<thead>
<tr>
<th></th>
<th>GALE-A</th>
<th>GALE-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMEANT</td>
<td>0.53</td>
<td>0.37</td>
</tr>
<tr>
<td>IMEANT</td>
<td><strong>0.51</strong></td>
<td><strong>0.33</strong></td>
</tr>
<tr>
<td>XMEANT</td>
<td>0.51</td>
<td>0.20</td>
</tr>
<tr>
<td>MEANT</td>
<td>0.48</td>
<td>0.33</td>
</tr>
<tr>
<td>METEOR 1.5 (2014)</td>
<td>0.43</td>
<td>0.10</td>
</tr>
<tr>
<td>NIST</td>
<td>0.29</td>
<td>0.16</td>
</tr>
<tr>
<td>METEOR 0.4.3 (2005)</td>
<td>0.20</td>
<td>0.29</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.20</td>
<td>0.27</td>
</tr>
<tr>
<td>TER</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>PER</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td>CDER</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>WER</td>
<td>0.10</td>
<td>0.26</td>
</tr>
</tbody>
</table>
IMEANT outperforms cross-lingual XMEANT

- IMEANT is tied with XMEANT on GALE-A
- IMEANT correlates with HAJ much better than XMEANT on GALE-B

| Table 1. Sent-level correlation with HAJ on GALE P2.5 data |
|---------------------------------|-----------------|-----------------|
|                                 | GALE-A          | GALE-B          |
| HMEANT                          | 0.53            | 0.37            |
| 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            |
IMEANT outperforms any of the others

- IMEANT produces much higher HAJ correlations than any of the other metrics on both GALE-A and GALE-B

<table>
<thead>
<tr>
<th>Table 1: Sent-level correlation with HAJ on GALE P2.5 data</th>
<th>GALE-A</th>
<th>GALE-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMEANT</td>
<td>0.53</td>
<td>0.37</td>
</tr>
<tr>
<td>IMEANT</td>
<td>0.51</td>
<td>0.33</td>
</tr>
<tr>
<td>XMEANT</td>
<td>0.51</td>
<td>0.20</td>
</tr>
<tr>
<td>MEANT</td>
<td>0.48</td>
<td>0.33</td>
</tr>
<tr>
<td>METEOR 1.5 (2014)</td>
<td>0.43</td>
<td>0.10</td>
</tr>
<tr>
<td>NIST</td>
<td>0.29</td>
<td>0.16</td>
</tr>
<tr>
<td>METEOR 0.4.3 (2005)</td>
<td>0.20</td>
<td>0.29</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.20</td>
<td>0.27</td>
</tr>
<tr>
<td>TER</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>PER</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td>CDER</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>WER</td>
<td>0.10</td>
<td>0.26</td>
</tr>
</tbody>
</table>
IMEANT even closes the gap with HMEANT

IMEANT even comes within a few points of the human upper bound established by HMEANT

Table 1. Sent-level correlation with HAJ on GALE P2.5 data

<table>
<thead>
<tr>
<th></th>
<th>GALE-A</th>
<th>GALE-B</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMEANT</td>
<td>0.53</td>
<td>0.37</td>
</tr>
<tr>
<td>IMEANT</td>
<td><strong>0.51</strong></td>
<td><strong>0.33</strong></td>
</tr>
<tr>
<td>XMEANT</td>
<td>0.51</td>
<td>0.20</td>
</tr>
<tr>
<td>MEANT</td>
<td>0.48</td>
<td><strong>0.33</strong></td>
</tr>
<tr>
<td>METEOR 1.5 (2014)</td>
<td>0.43</td>
<td>0.10</td>
</tr>
<tr>
<td>NIST</td>
<td>0.29</td>
<td>0.16</td>
</tr>
<tr>
<td>METEOR 0.4.3 (2005)</td>
<td>0.20</td>
<td>0.29</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.20</td>
<td>0.27</td>
</tr>
<tr>
<td>TER</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>PER</td>
<td>0.20</td>
<td>0.18</td>
</tr>
<tr>
<td>CDER</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>WER</td>
<td>0.10</td>
<td>0.26</td>
</tr>
</tbody>
</table>
observation

how ITG constraints help IMEANT

- empirically, we see
  - ITGs produce significantly more accurate phrasal similarity aggregation
  - compared to MEANT’s standard bag-of-words based heuristics

- permutation and bijectivity constraints enforced by the ITG
  - offer better leverage to reject inappropriate token alignments
  - compared to the maximal alignment approach which tends to be rather promiscuous
example

how ITG constraints help IMEANT

- clean, sparse alignments for the role fillers of ARG1 of the "resumed" PRED
- leaving tokens like "complete" and "range" unaligned (instead of aligning them anyway as MEANT’s maximal alignment does)

[MT2] So far, in the mainland of China to stop selling nearly two months of SK-II products sales resumed.

[REF] Until after their sales had ceased in mainland China for almost two months, sales of the complete range of SK-II products have now been resumed.
simple  Occam’s razor: easy to define, easy to implement, easy to use

representationally transparent  can look at a score and understand scientifically why it was high or low
  - eg, MEANT’s degree of match between semantic frames
  - who did what to whom, for whom, when, where, why and how

tunable  support fast scoring of massive numbers of hypotheses for tuning/training

discriminating  fine-grained scores (not just ranking or “good/bad” binary classification)

language independent  methodology that works across all language pairs
  - eg, IMEANT and XMEANT’s incorporation of language universal ITG biases

stable  high HAJ correlations without retraining
conclusions

- **IMEANT** – our newest 2014 version of MEANT is based on ITGs

- **achieves highest correlation** with HAJ among all variants of MEANT as well as other common MT evaluation metrics

- aligns and scores semantic frames via a simple, consistent BITG which provides **informative permutation and bijectivity biases**
  - replaces MEANT’s maximal alignment and bag-of-words heuristics

- retains MEANT’s characteristics of **Occam’s Razor style simplicity** and **representational transparency**