Unsupervised vs. Supervised Weight Estimation for Semantic MT Evaluation Metrics

Chi-kiu Lo    Dekai Wu

HKUST Human Language Technology Center
Department of Computer Science and Engineering
University of Science and Technology, Hong Kong

{ jackielo | dekai }@cs.ust.hk
The problem with conventional MT evaluation metrics

This has been our SMT trajectory over the years

- **1993-1995** First unstructured SMT on very different langs (Chinese)
- **1995-now** First syntactic SMT (ITG, BITG, phrasal ITG)
- **2009-now** Recent syntactic SMT (LTG, LITG, PLITG)
- **2005-now** First semantic SMT with WSD-for-SMT (PSD)
- **2007-now** First semantic SMT with SRL-for-SMT

Subjective evaluation shows improvement...
But conventional metrics like BLEU aren’t discriminating enough to register it

Serious danger of driving our field astray!

- **2009-now** Semantic MT evaluation with SRL-for-MTE (MEANT)
**Background**

*Acknowledgments: DARPA GALE, BOLT*

- **LREC 2010, SSST 2010**
  - Blueprint HMEANT model, preliminary results
- **ACL 2011**
  - Assesses adequacy via Propbank-style semantic predicates, roles, and fillers
  - Explains MT accuracy with high representational transparency
  - Correlates with human adequacy judgments (HAJ) as well as HTER, BUT at lower cost
- **IJCAI 2011**
  - “Flattened” HMEANT improves correlation with HAJ, by ignoring which frames roles/fillers are associated with (!!!)
  - Correlation of individual roles against HAJ
  - Analysis of time cost of evaluation
- **SSST 2011**
  - Back to compositionality – “unflattens” HMEANT and further improves correlation with HAJ
  - Weights the degree of contribution of each frame, according to size of the span it covers

Hong Kong University of Science & Technology
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! 😞

[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 .

So far, the sale in the mainland of China for nearly two months of SK-II line of products .

<table>
<thead>
<tr>
<th>N-gram</th>
<th>Syntax-subtree</th>
<th>SRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram matches: 15</td>
<td>1-level subtree matches: 34</td>
<td>Predicate matches: 0</td>
</tr>
<tr>
<td>2-gram matches: 4</td>
<td>2-level subtree matches: 8</td>
<td></td>
</tr>
<tr>
<td>3-gram matches: 3</td>
<td>3-level subtree matches: 2</td>
<td></td>
</tr>
<tr>
<td>4-gram matches: 1</td>
<td>4-level subtree matches: 0</td>
<td></td>
</tr>
</tbody>
</table>
Conversely: a more useful translation
More SRL matches 😊
but fewer N-gram and syntax-subtree matches! ☹

<table>
<thead>
<tr>
<th>N-gram</th>
<th>Syntax-subtree</th>
<th>SRL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-gram matches:</td>
<td>15 1-level subtree matches: 35</td>
<td>Predicate matches: 2</td>
</tr>
<tr>
<td>2-gram matches:</td>
<td>4 2-level subtree matches: 6</td>
<td>Argument matches: 1</td>
</tr>
<tr>
<td>3-gram matches:</td>
<td>1 3-level subtree matches: 1</td>
<td></td>
</tr>
<tr>
<td>4-gram matches:</td>
<td>0 4-level subtree matches: 0</td>
<td></td>
</tr>
</tbody>
</table>
**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
HMEANT is fairly cheap...
...but still requires humans

- **Annotation tasks**
  1. label semantic predicates, roles, and fillers
  2. align predicates and fillers between the reference and machine translations

- **Ranking task**
  - label human adequacy judgment to form a training corpus for the role importance
Unsupervised weight estimates are needed

- Testing HMEANT on WMT-2012 English-Czech (w/ Bojar et al.)
  - Manpower constraint: 14 Czech-speaking annotators
  - Time constraint: within two days
  - Translation of 50 sentences from 13 systems and 1 reference

- What about the labeled training data?
  - No more resources (Czech speakers)
  - Applying the weights learned from English data is obscured
    - linguistic differences between Czech and English,
      e.g. dropping of pronoun in Czech

Hong Kong University of Science & Technology
Our goal:

- **Further reduce the cost of evaluating MT** by eliminating the dependency on a human adequacy-ranked training corpus for tuning the weights for each semantic role type.

- Here, we’re mainly targeting the problem of evaluating translation quality for languages with sparse resources.
Using relative frequency to estimate MEANT’s parameters

- Basic assumption:
  - Roles that are more important for humans to understand should appear more often in the language

- We propose an **unsupervised** approach:
  - Use the relative frequency of how often a type of semantic role appears in reference translations, to estimate the degree of contribution of that role type

\[
c_j \equiv \text{# count of ARG } j \text{ in REF of the test set}
\]

\[
w_j = \frac{c_j}{\sum_j c_j}
\]

Hong Kong University of Science & Technology
Correctness of the proposed unsupervised approach

- Problem: No ground truth on which role type contributes more to the overall meaning

- Solution: Evaluate how closely the unsupervised weight of each role type approximates the weight obtained from supervised training
Results

- Relative frequency of each semantic role type closely approximates the supervised weight of that type

<table>
<thead>
<tr>
<th>Role</th>
<th>Deviation (GALE-A)</th>
<th>Deviation (GALE-B)</th>
<th>Deviation (WMT12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent</td>
<td>-0.09</td>
<td>-0.05</td>
<td>0.03</td>
</tr>
<tr>
<td>Experiencer</td>
<td>0.23</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>Benefactive</td>
<td>0.02</td>
<td>0.04</td>
<td>-0.01</td>
</tr>
<tr>
<td>Temporal</td>
<td>0.11</td>
<td>0.08</td>
<td>0.03</td>
</tr>
<tr>
<td>Locative</td>
<td>-0.05</td>
<td>-0.05</td>
<td>-0.07</td>
</tr>
<tr>
<td>Purpose</td>
<td>-0.01</td>
<td>0.03</td>
<td>-0.01</td>
</tr>
<tr>
<td>Manner</td>
<td>-0.01</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td>Extent</td>
<td>-0.02</td>
<td>0.00</td>
<td>-0.01</td>
</tr>
<tr>
<td>Modal</td>
<td>—</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>Negation</td>
<td>—</td>
<td>0.01</td>
<td>-0.01</td>
</tr>
<tr>
<td>Other</td>
<td>-0.12</td>
<td>0.05</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

Table 1: Deviation of relative frequency from optimized weight of each semantic role in GALE-A, GALE-B and WMT12

Hong Kong University of Science & Technology
Estimating the weight for the predicate

- Treating predicate the same way as the arguments
  - Using relative frequency of the predicate in addition to all semantic arguments
    
    \[
    c_{\text{pred}} \equiv \# \text{ count of PRED in REF of the test set}
    \]
    
    Method (i) \[= \frac{c_{\text{pred}}}{c_{\text{pred}} + \sum_j c_j}\]

- BUT, predicates are fundamentally different from arguments
  - Every semantic is defined by one predicate, and arguments are defined relative to the predicate

- In the supervised weights, predicate is usually one-fourth as important as the agent role
  
  \[
  \text{Method (ii)} = 0.25 \cdot w_{\text{agent}}
  \]
Results

- The heuristic of one-fourth of the agent’s weight closely approximates the weight of the predicate

<table>
<thead>
<tr>
<th>PRED estimation</th>
<th>Deviation (GALE-A)</th>
<th>Deviation (GALE-B)</th>
<th>Deviation (WMT12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method (i)</td>
<td>0.16</td>
<td>0.16</td>
<td>0.31</td>
</tr>
<tr>
<td>Method (ii)</td>
<td>0.02</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 2: Deviation from optimized weight in GALE-A, GALE-B and WMT12 of the predicate’s weight as estimated by (i) frequency of predicates in frames, relative to predicates and arguments; and (ii) one-fourth of agent’s weight.
HMEANT using unsupervised weight estimates

- Unsupervised approach closely approximates the weights obtained from supervised approach

- Then, comparing to other MT evaluation metrics, how does HMEANT using unsupervised weights perform?
Results

- Unsupervised HMEANT correlates with HAJ comparably to supervised HMEANT

<table>
<thead>
<tr>
<th>Metrics</th>
<th>GALE-A</th>
<th>GALE-B</th>
<th>WMT12</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMEANT (supervised)</td>
<td>0.49</td>
<td>0.27</td>
<td>0.29</td>
</tr>
<tr>
<td>HMEANT (unsupervised)</td>
<td>0.42</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td>NIST</td>
<td>0.29</td>
<td>0.09</td>
<td>0.12</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.20</td>
<td>0.21</td>
<td>0.22</td>
</tr>
<tr>
<td>TER</td>
<td>0.20</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>PER</td>
<td>0.20</td>
<td>0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.20</td>
<td>0.12</td>
<td>0.01</td>
</tr>
<tr>
<td>CDER</td>
<td>0.12</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>WER</td>
<td>0.10</td>
<td>0.11</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table 3: Average sentence-level correlation with human adequacy judgments of HMEANT using supervised and unsupervised weight scheme on GALE-A, GALE-B and WMT12, (with baseline comparison of commonly used automatic MT evaluation metric.)
Conclusion

- Using relative frequency of semantic roles (unsupervised) to estimate HMEANT’s parameters:
  - **further reduces the evaluation cost** by eliminating the dependency on a human adequacy-ranked training corpus for tuning the weights for each semantic role type
  - **correlates with HAJ** comparably to supervised HMEANT on all three data set, including WMT-2012 English-Czech
  - **is well suited to sparse languages** for evaluating translation
Progress toward automating HMEANT...

- Fully automated MEANT (WMT-2012, at NAACL, in June 2012)
  - First fully automated semantic MT evaluation metric
  - Replaces human SRL with automatic shallow semantic parsing
  - Replaces human semantic frame alignment with a simple maximum weighted bipartite matching algorithm based on the lexical similarity between semantic frames
  - Preserves the spirit of Occam’s razor of HMEANT
  - Outperforms all commonly used automatic metrics

- Training SMT with MEANT as the objective function
  - Minimum error rate training runs completed two weeks ago
  - Highly competitive results
  - In progress: Human quality evaluation on MT output tuned on MEANT vs. BLEU vs. TER