Towards a Predicate-Argument Evaluation for MT

Ondřej Bojar\textsuperscript{\textalpha}, Dekai Wu\textsuperscript{\textbeta}

\textsuperscript{\textalpha} Charles University in Prague, ÚFAL
\textsuperscript{\textbeta} Hong Kong University of Science and Technology, HKUST

bojar@ufal.mff.cuni.cz, dekai@cs.ust.hk

presented by Rudolf Rosa

SSST-6 July 12, 2012
Outline

- Motivating predicate-argument evaluation for MT.
  - HMEANT Annotation Procedure.
  - Relation to Tectogrammatical Annotation.
- Experiment with Czech.
- Issues Encountered.
- Summary.
Methods of Manual MT Evaluation

- Absolute **adequacy** and **fluency** of whole sentences.
- **Ranking of full sentences.**
- **Ranking of constituents**, i.e. parts of sentences.
- **Comprehension test:** Blind editing + validation.
- **Task-based:** MT output as useful as the original? Do I dress appropriately given a translated weather forecast?
- **HTER:** Post-editing effort.
Methods of Manual MT Evaluation

- Absolute **adequacy** and **fluency** of whole sentences. Measures correlated. Low agreement.

- Ranking of full sentences.

- Ranking of constituents, i.e. parts of sentences.

- **Comprehension test**: Blind editing + validation.

- **Task-based**: MT output as useful as the original? Do I dress appropriately given a translated weather forecast?

- **HTER**: Post-editing effort.
Methods of Manual MT Evaluation

- **Absolute adequacy** and **fluency** of whole sentences. Measures correlated. Low agreement.


- **Ranking of constituents**, i.e. parts of sentences.

- **Comprehension test:** Blind editing+validation.

- **Task-based**: MT output as useful as the original? Do I dress appropriately given a translated weather forecast?

- **HTER**: Post-editing effort.
Methods of Manual MT Evaluation

- **Absolute adequacy and fluency** of whole sentences. Measures correlated. Low agreement.

- **Ranking of full sentences.**
  Longer sentences hard to rank. Candidates incomparably poor.

- **Ranking of constituents**, i.e. parts of sentences.
  Does not evaluate overall coherence.

- **Comprehension test**: Blind editing + validation.

- **Task-based**: MT output as useful as the original?
  Do I dress appropriately given a translated weather forecast?

- **HTER**: Post-editing effort.
Methods of Manual MT Evaluation

- **Absolute adequacy** and **fluency** of whole sentences. Measures correlated. Low agreement.


- **Ranking of constituents**, i.e. parts of sentences. Does not evaluate overall coherence.

- **Comprehension test:** Blind editing + validation. Expensive.

- **Task-based:** MT output as useful as the original? Do I dress appropriately given a translated weather forecast?

- **HTER:** Post-editing effort.
Methods of Manual MT Evaluation

- **Absolute adequacy** and **fluency** of whole sentences. Measures correlated. Low agreement.

- **Ranking of full sentences.**
  Longer sentences hard to rank. Candidates incomparably poor.

- **Ranking of constituents**, i.e. parts of sentences.
  Does not evaluate overall coherence.

- **Comprehension test**: Blind editing + validation.
  Expensive.

- **Task-based**: MT output as useful as the original?
  Do I dress appropriately given a translated weather forecast?
  Preparation expensive. Feels too narrow.

- **HTER**: Post-editing effort.
Methods of Manual MT Evaluation

- **Absolute adequacy** and **fluency** of whole sentences. Measures correlated. Low agreement.
- **Ranking of full sentences.**
  Longer sentences hard to rank. Candidates incomparably poor.
- **Ranking of constituents**, i.e. parts of sentences.
  Does not evaluate overall coherence.
- **Comprehension test**: Blind editing + validation. Expensive.
- **Task-based**: MT output as useful as the original?
  Do I dress appropriately given a translated weather forecast?
  Preparation expensive. Feels too narrow.
- **HTER**: Post-editing effort.
  Expensive. Requires trained people.
HMEANT (Lo and Wu, 2011a)

- Improved evaluation of adequacy compared to BLEU.
- Reduced human labour of HTER (Snover et al., 2006).

Essence: Is the basic event structure understandable? 
Who did what to whom, when, where and why.

Procedure:
1. SRL: Identify semantic frames and roles in ref & hyp.
2. Align frames and their role fillers.
3. Calculate prec & rec across all frames in the sentence.
HMEANT Illustration: Motivation

It is hard to rank A vs. B (even if we know R is the ref.)

A Finally, he stood in the center of the referee Wolfgang Stark.

B At the end of the day, was at the centre of a referee Wolfgang Stark.

R The referee Wolfgang Stark then garnered some attention.
Finally, he stood in the center of the referee Wolfgang Stark.

It is easier to mark roles of a single hypothesis.
It is easier to mark roles of a single hypothesis.

Finally, he **stood** in the center of the referee Wolfgang Stark.
It is easier to mark roles of a single hypothesis.

Finally, he stood in the center of the referee Wolfgang Stark.
Finally, he stood in the center of the referee Wolfgang Stark.

It is easier to mark roles of a single hypothesis.
The same SRL is performed on the reference.

The referee Wolfgang Stark then garnered some attention.
The same SRL is performed on the reference.

The referee Wolfgang Stark then garnered some attention.
Finally, he stood in the center of the referee Wolfgang Stark.

At the end of the day, was at the center of a referee Wolfgang Stark.

The referee Wolfgang Stark then garnered some attention.

It is hard to rank A vs. B (even if we know R is the ref.)

The same SRL is performed on the reference.
And finally, frames and role fillers are aligned.

Finally, he stood in the center of the referee Wolfgang Stark.

The referee Wolfgang Stark then garnered some attention.
And finally, frames and role fillers are aligned.

A: Finally, he stood in the center of the referee Wolfgang Stark.

R: The referee Wolfgang Stark then garnered some attention.
And finally, frames and role fillers are aligned.

A
Finally, he stood in the center of the referee Wolfgang Stark.

R
The referee Wolfgang Stark then garnered some attention.

Obviously, the meaning was rather distorted.
And finally, frames and role fillers are aligned.

Finally, he stood in the center of the referee Wolfgang Stark.

...but the annotation was more principled and we know which parts are wrong.

The referee Wolfgang Stark then garnered some attention.
Finally, he stood in the center of the referee Wolfgang Stark.

In terms of pred-arg. formalisms like the tectogrammatical layer of the Prague Dependency Treebank (PDT):

- HMEANT just checks the match of subtrees under verbs.
- Tools for English and Czech available to get such trees automatically.

⇒ We could e.g. highlight all words of a subtree at once.
English→Czech Experiment

- 50 distinct sentences from WMT12 test set.
  - Selected to have a high overlap with WMT12 manual rankings for future analysis.

- 13 systems translating from English to Czech.
  + One reference translation.

- 14 annotators
  - No sentence displayed twice to the same person.

- Unfortunately no overlap in annotation
  ⇒ No agreement judgments.
Sentence-Level Correlation

<table>
<thead>
<tr>
<th>Metric</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMEANT</td>
<td>0.2833</td>
</tr>
<tr>
<td>METEOR</td>
<td>0.2167</td>
</tr>
<tr>
<td>WER</td>
<td>0.1708</td>
</tr>
<tr>
<td>CDER</td>
<td>0.1375</td>
</tr>
<tr>
<td>NIST</td>
<td>0.1167</td>
</tr>
<tr>
<td>TER</td>
<td>0.1167</td>
</tr>
<tr>
<td>PER</td>
<td>0.0208</td>
</tr>
<tr>
<td>BLEU</td>
<td>0.0125</td>
</tr>
</tbody>
</table>

- Better correlation than automatic metrics (expected).
- Overall quite low. Possible reasons:
  - Evaluated 13 systems.
  - Gold standard ranks overall quality, not just adequacy as Lo and Wu (2011b) who achieve 0.49.
  - HMEANT problems discovered by our experiment, see below.
  - Gold standard disputable.

Kendall’s $\tau$ for sentence-level correlation with human rankings.
## Gold Standard

<table>
<thead>
<tr>
<th>Interpretation Sentences</th>
<th>Ties Ignored</th>
</tr>
</thead>
<tbody>
<tr>
<td>cu-depfix</td>
<td>72.5</td>
</tr>
<tr>
<td>onlineB</td>
<td>61.4</td>
</tr>
<tr>
<td>uedin-wmt12</td>
<td>60.3</td>
</tr>
<tr>
<td>cu-tamch-boj</td>
<td>54.6</td>
</tr>
<tr>
<td>cu-bojar_2012</td>
<td>53.2</td>
</tr>
<tr>
<td>CU_TectoMT</td>
<td>54.9</td>
</tr>
<tr>
<td>onlineA</td>
<td>61.4</td>
</tr>
<tr>
<td>pctrans2010</td>
<td>54.1</td>
</tr>
<tr>
<td>commercial2</td>
<td>51.3</td>
</tr>
<tr>
<td>cu-poor-comb</td>
<td>41.6</td>
</tr>
<tr>
<td>uk-dan-moses</td>
<td>33.2</td>
</tr>
<tr>
<td>SFU</td>
<td>31.0</td>
</tr>
<tr>
<td>jhu-hiero</td>
<td>26.7</td>
</tr>
</tbody>
</table>
## Somewhat Shaky Gold Standard

<table>
<thead>
<tr>
<th>Interpretation</th>
<th>Ties Ignored</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentences</td>
<td>All</td>
</tr>
<tr>
<td>cu-depfix</td>
<td>66.4</td>
</tr>
<tr>
<td>onlineB</td>
<td>63.0</td>
</tr>
<tr>
<td>uedin-wmt12</td>
<td>55.8</td>
</tr>
<tr>
<td>cu-tamch-boj</td>
<td>55.6</td>
</tr>
<tr>
<td>cu-bojar_2012</td>
<td>54.3</td>
</tr>
<tr>
<td>CU_TectoMT</td>
<td>53.1</td>
</tr>
<tr>
<td>onlineA</td>
<td>52.9</td>
</tr>
<tr>
<td>pctrans2010</td>
<td>47.7</td>
</tr>
<tr>
<td>commercial2</td>
<td>46.0</td>
</tr>
<tr>
<td>cu-poor-comb</td>
<td>44.1</td>
</tr>
<tr>
<td>uk-dan-moses</td>
<td>43.5</td>
</tr>
<tr>
<td>SFU</td>
<td>36.1</td>
</tr>
<tr>
<td>jhu-hiero</td>
<td>32.2</td>
</tr>
</tbody>
</table>
## Somewhat Shaky Gold Standard

<table>
<thead>
<tr>
<th>Interpretation Sentences</th>
<th>Ties Ignored</th>
<th>≥ Others</th>
<th>&gt; Others</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>50</td>
<td>All</td>
</tr>
<tr>
<td>cu-depfix</td>
<td>66.4</td>
<td>72.5</td>
<td>73.0</td>
</tr>
<tr>
<td>onlineB</td>
<td>63.0</td>
<td>61.4</td>
<td>70.5</td>
</tr>
<tr>
<td>uedin-wmt12</td>
<td>55.8</td>
<td>60.3</td>
<td>63.6</td>
</tr>
<tr>
<td>cu-tamch-boj</td>
<td>55.6</td>
<td>54.6</td>
<td>64.7</td>
</tr>
<tr>
<td>cu-bojar_2012</td>
<td>54.3</td>
<td>53.2</td>
<td>64.1</td>
</tr>
<tr>
<td>CU_TectoMT</td>
<td>53.1</td>
<td>54.9</td>
<td>60.5</td>
</tr>
<tr>
<td>onlineA</td>
<td>52.9</td>
<td>61.4</td>
<td>60.8</td>
</tr>
<tr>
<td>pctrans2010</td>
<td>47.7</td>
<td>54.1</td>
<td>55.1</td>
</tr>
<tr>
<td>commercial2</td>
<td>46.0</td>
<td>51.3</td>
<td>54.6</td>
</tr>
<tr>
<td>cu-poor-comb</td>
<td>44.1</td>
<td>41.6</td>
<td>54.7</td>
</tr>
<tr>
<td>uk-dan-moses</td>
<td>43.5</td>
<td>33.2</td>
<td>53.4</td>
</tr>
<tr>
<td>SFU</td>
<td>36.1</td>
<td>31.0</td>
<td>46.8</td>
</tr>
<tr>
<td>jhu-hiero</td>
<td>32.2</td>
<td>26.7</td>
<td>43.2</td>
</tr>
</tbody>
</table>
Problems of HMEANT Annotation

- Vague SRL Guidelines:
  - Complex predicates.
  - PP-attachment.
  - Unclear or insufficient role labels.
  - Co-reference.

- Problems in the Alignment Phase:
  - Correctness of the Predicate.
  - Need for M:N Frame and Slot Alignment.
Complex Predicates

HMEANT tool requires exactly 1 word to serve as Action.

- Modals have a separate “role” label.

In Czech:

- It is the modal that conjugates ⇒ disputable.

<table>
<thead>
<tr>
<th>Czech (made up)</th>
<th>Představení</th>
<th>musí</th>
<th>pokračovat.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloss</td>
<td>Show</td>
<td>must</td>
<td>go on.</td>
</tr>
<tr>
<td>English-Like Labels</td>
<td>Agent</td>
<td>Modal</td>
<td>Action</td>
</tr>
<tr>
<td>Natural for Czech</td>
<td>Agent</td>
<td>Action</td>
<td>Action</td>
</tr>
<tr>
<td>Forced to 1 Word</td>
<td>Agent</td>
<td>Action</td>
<td>Experiencer</td>
</tr>
</tbody>
</table>

- Copula “to be” is frequent.

<table>
<thead>
<tr>
<th>Czech (made up)</th>
<th>Řidič</th>
<th>byl</th>
<th>unaven.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloss</td>
<td>The driver</td>
<td>was</td>
<td>tired.</td>
</tr>
<tr>
<td>English-Like Labels</td>
<td>Agent</td>
<td>Action</td>
<td>Experiencer?</td>
</tr>
<tr>
<td>Natural for Czech</td>
<td>Agent</td>
<td>Action</td>
<td>Action</td>
</tr>
</tbody>
</table>

⇒ We suggest allowing more words to denote an Action.
Prepositional Phrase (PP) Attachment

<table>
<thead>
<tr>
<th>Reference</th>
<th>Oblečky</th>
<th>musíme</th>
<th>vystříhat</th>
<th>z časopisů</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloss</td>
<td>clothes</td>
<td>we-must</td>
<td>cut</td>
<td>from magazines</td>
</tr>
<tr>
<td>Roles</td>
<td>Experiencer</td>
<td>Modal</td>
<td>Action</td>
<td>Locative</td>
</tr>
<tr>
<td>Meaning</td>
<td>We must cut the clothes (paper toys) from magazines</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Musíme</th>
<th>vyříznout</th>
<th>oblečení z časopisů</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gloss</td>
<td>We-must</td>
<td>cut</td>
<td>clothes from magazines</td>
</tr>
<tr>
<td>Roles</td>
<td>Modal</td>
<td>Action</td>
<td>Experiencer</td>
</tr>
</tbody>
</table>

- The PP “from magazines” in the hypothesis can be annotated as:
  - a separate Locative.
  - or a part of Experiencer.
  (Sometimes the separate annotation is forced by word order.)

⇒ Impossible to align 2 to 1 role fillers.
⇒ Translation quality underestimated.
Unclear or Insufficient Role Labels

- HMEANT requires role labels to match to give credit.
- HMEANT set of labels is sufficiently simple:
  - So the disagreement is hopefully kept low.
  - Sometimes still hard to use, e.g. in passive constructions. ⇒ Disagreement ⇒ Translation quality underestimated.

- On the other hand, the set feels too small in some cases:
  - **Czech** Byl převezen do nemocnice ve vrtulníku.
  - **Gloss** He was transported to the hospital in a helicopter.
  - **SRL** Action Locative Locative

  - One of out annots. actually joined the two Locatives into one. ⇒ 2:1 alignment problem.

⇒ We suggest experimenting with no role labels altogether.
Co-reference

Consider annotation of the frame of “wins”:

<table>
<thead>
<tr>
<th>English (made up)</th>
<th>It</th>
<th>is</th>
<th>the man</th>
<th>who</th>
<th>wins.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roles</td>
<td>Agent?</td>
<td>Agent?</td>
<td>Agent?</td>
<td>Action</td>
<td></td>
</tr>
</tbody>
</table>

- Three candidates for the Agent.
- In Czech, some can be pro-dropped.
  - Risk of no Agent annotated at all.
  - 0:1-alignment problem.
  - Translation quality underestimated.

⇒ We suggest giving more examples in the guidelines.
Correctness of the Predicate

| Reference | Opilý řidič | těžce | zraněn | Gloss: A drunken driver seriously injured (passive form) |
| Roles     | Agent      | Extent| Action | |
| Meaning   | A drunken driver is seriously injured. | |

| Hypothesis | Opilý řidič | vážně | zranil | Gloss: A drunken driver seriously injured (active form) |
| Roles      | Agent      | Extent| Action | |
| Meaning    | A drunken driver seriously injured (someone). | |

- All role fillers match exactly.
- The **Action**’s form reverses the meaning.
- Current HMEANT does not allow to mark Action as mistranslated.

⇒ We suggest judging the quality of **predicate match** as well.
Need for M:N Frame and Slot Alignments

HMEANT aligns first frames and then slots within them.

- But the frames do not always match 1-1, e.g. due to:
  - inconsistent annotation of modals, phasic verbs ("to begin")
  - or simply not quite literal but correct translation.

⇒ Cannot align fillers across frames.
⇒ Translation quality underestimated.

PP-attachment ambiguity:

- Happens in the SRL phase.
- Causes a 2:1 problem in the alignment phase.
⇒ Translation quality underestimated.
⇒ We suggest allowing M:N ali. for both frames and fillers.
Summary

We applied HMEANT to Czech.

- Overall positive experience.
  - Annotators know what they are doing when, where and why.
- Multiple issues identified:
  - Some can be solved by more examples to current guidelines.
  - Some require an update of the interface.
    - Multiple (non-adjacent) words forming the Action.
    - Indication of the correctness of the predicate.
  - Some need changes to prec/rec formulas.
    - M:N alignments of predicates and slots.

Future: Use t-layer tools to:

- Speed up SRL (highlight more words at once).
- Fully automate HMEANT \(\rightsquigarrow\) MEANT.
