Learning an Expert from Human Annotations in Statistical Machine Translation: the Case of Out-of-Vocabulary Words

Wilker Aziz\textsuperscript{(1)}, Marc Dymetman\textsuperscript{(2)}, Shachar Mirkin\textsuperscript{(3)}, Lucia Specia\textsuperscript{(4)}, Nicola Cancetta\textsuperscript{(2)} and Ido Dagan\textsuperscript{(3)}

\textsuperscript{(1)} São Paulo University \hspace{1cm} \textsuperscript{(2)} Xerox Research Centre Europe; \\
\textsuperscript{(3)} Bar Ilan University \hspace{1cm} \textsuperscript{(4)} University of Wolverhampton
The mayor was *attacked* by the press
Le maire a été *attacked* par la presse

(mayor, maire)
(press, presse)
(attacked, ?)
OOV words and paraphrase/entailment

The mayor was *attacked* by the press
Le maire a été *attacked* par la presse

entailments

attacked → accused
attacked → hit

(call, *accusé*)
(hit, touché)

Le maire a été *accusé* par la presse

Callison-Burch et al, 2006; Marton et al, 2009: **paraphrases**
Mirkin et al, 2009: **entailments**
OOV entailments as a learning problem

- By contrast to previous work on paraphrase/entailments for OOV in SMT, we cast replacement selection as:
  - a learning problem
  - from human annotations
  - with the entailment model tightly integrated into the Phrase-Based SMT decoder

- Learning an OOV expert for a PB-SMT system
Learning an expert for OOV sentences

• Integrated PB-SMT model, that includes an expert for OOV sentences
  – One overall SMT model, built on top of standard PB-SMT model
  – Contextual features, representing properties of the replacements
  – “Dynamic” biphrases built on demand

• Learning from human judgments
  – Annotators rank translations corresponding to different replacement choices
  – The integrated SMT model is tuned in order to bring system ranking close to annotator ranking

• Active learning:
  – For each OOV sentence, only a few candidate translations are shown, depending on current state of the model

• Avoiding to bias the SMT system towards OOV sentences
  – Learning is done in such a way that the integrated model behaves like the standard model on standard sentences
Dynamic biphrases

Source entailment

attacked $\rightarrow$ accused
attacked $\rightarrow$ hit

Static biphrase

(accused, accusé)
(hit, touché)

Dynamic biphrase

(attacked, accusé)
(attacked, touché)
# Features

## Entailment features

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>$DSim$</th>
<th>$CSim$</th>
<th>InfoLoss</th>
</tr>
</thead>
<tbody>
<tr>
<td>attacked → accused</td>
<td>-3.1</td>
<td>-0.3</td>
<td>-0.4</td>
<td></td>
</tr>
<tr>
<td>attacked → hit</td>
<td>-5.2</td>
<td>-7.2</td>
<td>-0.5</td>
<td></td>
</tr>
</tbody>
</table>

## Biphrase features

<table>
<thead>
<tr>
<th>Source</th>
<th>Target</th>
<th>$DSim$</th>
<th>$CSim$</th>
<th>InfoLoss</th>
<th>Clone</th>
</tr>
</thead>
<tbody>
<tr>
<td>static biphrases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mayor</td>
<td>maire</td>
<td>-0.1</td>
<td>-0.1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>press</td>
<td>presse</td>
<td>-1.5</td>
<td>-0.7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>accused</td>
<td>accusé</td>
<td>-1.6</td>
<td>-1.2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>hit</td>
<td>touché</td>
<td>-0.9</td>
<td>-0.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>dynamic biphrases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>attacked</td>
<td>accusé</td>
<td>-3.1</td>
<td>-0.3</td>
<td>-0.4</td>
<td>-1.6</td>
</tr>
<tr>
<td>attacked</td>
<td>touché</td>
<td>-5.2</td>
<td>-7.2</td>
<td>-0.5</td>
<td>-0.9</td>
</tr>
</tbody>
</table>
Entailment features: details

- **Contextual score (CSIM):**
  - How well does *rep* fit the context of the sentence *s*
  - Based on cosine similarity of LSA vectors representing *rep* and *s*

- **Domain similarity (DSIM):**
  - How well does *rep* replaces *oov* in general texts of the domain
  - Based on cosine similarity of LSA vectors of *oov* and *rep*

- **Information Loss (InfoLoss):**
  - Measures distance between *oov* and *rep* in Wordnet

- **Other entailment features:**
  - Synonym/Hypernym (from Wordnet)
  - Identity replacement (replacement by copying source word)
The integrated model

• Original model

\[
\text{argmax}_{(a,t)} \Lambda \cdot G(s,t,a)
\]

G: standard "static" features

• Integrated model

\[
\text{argmax}_{(a,t)} \Lambda \cdot G(s,t,a) + M \cdot H(s,t,a)
\]

H: "dynamic" features

– Integrated feature vector: \( F = G \oplus H \)

– Integrated parameter vector: \( \Omega = \Lambda \oplus M \)
Given an initial value for $\Omega = \Lambda \oplus M$, and an OOV sentence $s$ ...

... actively sample around a dozen different translations for $s$ (out of many more candidates)

- According to probabilities assigned by $\Omega$ to these translations (but always include $\Omega$-best translation)
- But also including top candidates relative to individual features (contextual score, domain similarity, ...)
Human annotations: (2) Annotation interface

• Present these translations in an annotation interface
  – Ask annotators to concentrate on "closeness of meaning" for portions affected by the replacements
    • BLEU would be inadequate for this
  – Discourage too fine distinctions:
    • translations grouped in a few clusters
Human annotations:
(3) Update the parameters

• Update $\Omega$ from the annotation data:
  – Try to bring model rank and annotation rank closer
    • Whenever two translations $(s, t_j)$ and $(s, t_k)$ are ordered differently by the annotator and by the model
      ... then change $\Omega$ into $\Omega'$, in such a way that:
        1. $\Omega'$ now ranks $(s, t_j)$ and $(s, t_k)$ in the same order as the annotator
        2. $\Omega'$ moves from $\Omega$ as little as possible (in terms of Euclidian distance)
        3. If $\Omega = \Lambda \oplus M$, then $\Omega' = \Lambda \oplus M'$ (update does not change $\Lambda$)
Human annotations:
(3) Update the parameters

- Use $\Omega'$ for the next round of active sampling
  - For efficiency, $\Omega$ is only updated after batches of 80 source sentences
Experimental setup

• Baseline phrase-based SMT system: MATRAX
  – Trained on English-French Europarl data (1M sents)

• Training of integrated expert model:
  – 75,000 sents from WMT-09 News Commentary
  – Around 15% OOV sentences
  – Tuning set: 1,000 OOV sents
    • Two annotators
    • Active sampling on batches of 80 sents
    • Convergence of performance after 6 slices (480 sents)
  – Evaluation set: 500 OOV sents
    • Comparison of different systems
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>μ</th>
<th>σ</th>
<th>Best</th>
<th>Acceptance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert-Human'</td>
<td>2.274</td>
<td>1.803</td>
<td>0.6258</td>
<td>0.7002</td>
</tr>
<tr>
<td>Mirkin09-1</td>
<td>2.736</td>
<td>1.933</td>
<td>0.5172</td>
<td>0.5822</td>
</tr>
<tr>
<td>Mirkin09-2</td>
<td>2.744</td>
<td>1.931</td>
<td>0.5132</td>
<td>0.5822</td>
</tr>
<tr>
<td>Expert-Human</td>
<td>3.018</td>
<td>1.913</td>
<td>0.4145</td>
<td>0.5252</td>
</tr>
<tr>
<td>Expert-MERT</td>
<td>3.153</td>
<td>1.928</td>
<td>0.4024</td>
<td>0.4849</td>
</tr>
<tr>
<td>SMT-Baseline</td>
<td>3.998</td>
<td>1.603</td>
<td>0.1549</td>
<td>0.2918</td>
</tr>
<tr>
<td>Stat-Paraphrases</td>
<td>4.107</td>
<td>1.584</td>
<td>0.1690</td>
<td>0.2495</td>
</tr>
</tbody>
</table>

- **SMT-baseline**: The base SMT system MATRAX
- **Mirkin09-1,-2**: Two best `entailment’ systems from Mirkin et al, 2009: replacement choices not integrated in decoder, no training of the expert
- **Stat-Paraphrases**: An implementation of Marton-09, with new static biphrases obtained through paraphrases from original static biphrases
- **Expert-Human**: The model of this paper, trained from human annotations
- **Expert-Human’**: Identity replacements blocked at decoding time
- **Expert-Mert**: The model of this paper but trained by MERT
Conclusion

• OOV: an instance of a more general problem: Learning an Expert for SMT

On the basis of an existing SMT system

... And on a narrow domain of « expertise »

... Improve the performance of the system

... Based on human judgments for the narrow domain

... Without degrading the behavior of the system on sentences outside of the narrow domain
BACKUP
Thanks!
Learning to rank using MIRA

\[ \text{If } \Omega \cdot \Phi(y_{j,k}) \geq 0 \text{ then } \Omega' := \Omega \]
\[ \text{Else } \Omega' := \arg\min_{\omega} \|\omega - \Omega\|^2 \]
\[ \text{s.t. } \omega \cdot \Phi(y_{j,k}) - \omega \cdot \Phi(y_{k,j}) \geq 1 \]
\[ \text{and } \pi^{(\Lambda)}(\omega) = \Lambda_0 \]
# Feature combinations

<table>
<thead>
<tr>
<th>Features</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>Best</th>
<th>Acceptance</th>
</tr>
</thead>
<tbody>
<tr>
<td>LID</td>
<td>2.477</td>
<td>1.465</td>
<td>0.4728</td>
<td>0.5252</td>
</tr>
<tr>
<td>ID</td>
<td>2.491</td>
<td>1.463</td>
<td>0.4668</td>
<td>0.5211</td>
</tr>
<tr>
<td>LI</td>
<td>2.547</td>
<td>1.457</td>
<td>0.4427</td>
<td>0.5050</td>
</tr>
<tr>
<td>I</td>
<td>2.561</td>
<td>1.463</td>
<td>0.4447</td>
<td>0.4970</td>
</tr>
<tr>
<td>D</td>
<td>2.924</td>
<td>1.414</td>
<td>0.3360</td>
<td>0.3722</td>
</tr>
<tr>
<td>LD</td>
<td>2.930</td>
<td>1.412</td>
<td>0.3340</td>
<td>0.3702</td>
</tr>
<tr>
<td>L</td>
<td>3.056</td>
<td>1.361</td>
<td>0.2857</td>
<td>0.3300</td>
</tr>
<tr>
<td>Baseline</td>
<td>3.219</td>
<td>1.252</td>
<td>0.2093</td>
<td>0.2918</td>
</tr>
</tbody>
</table>

**Table 1:** Comparison between different feature combinations and the baseline showing the percentage of times each combination outputs a translation that is acceptable, i.e. is not discarded (Acceptance), a translation that is ranked in the first cluster (Best), as well as the the mean rank ($\mu$) and standard deviation ($\sigma$) of each combination, where the discarded translations are conventionally assigned a rank of 5, lower than the rank of any acceptable cluster observed among the annotations. (L) context model score, (I) information-loss, (D) domain similarity, (Baseline) SMT system.
Learning iterations improve results

<table>
<thead>
<tr>
<th>Iterations</th>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>Best</th>
<th>Acceptance</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_6$</td>
<td>2.487</td>
<td>1.458</td>
<td>0.4628</td>
<td>0.5252</td>
</tr>
<tr>
<td>$M_5$</td>
<td>2.491</td>
<td>1.459</td>
<td>0.4628</td>
<td>0.5231</td>
</tr>
<tr>
<td>$M_4$</td>
<td>2.489</td>
<td>1.458</td>
<td>0.4628</td>
<td>0.5252</td>
</tr>
<tr>
<td>$M_3$</td>
<td>2.493</td>
<td>1.455</td>
<td>0.4588</td>
<td>0.5252</td>
</tr>
<tr>
<td>$M_2$</td>
<td>2.501</td>
<td>1.456</td>
<td>0.4567</td>
<td>0.5211</td>
</tr>
<tr>
<td>$M_1$</td>
<td>2.519</td>
<td>1.456</td>
<td>0.4507</td>
<td>0.5151</td>
</tr>
<tr>
<td>$M_0$</td>
<td>2.944</td>
<td>1.407</td>
<td>0.328</td>
<td>0.3642</td>
</tr>
<tr>
<td>Baseline</td>
<td>3.237</td>
<td>1.228</td>
<td>0.1932</td>
<td>0.2918</td>
</tr>
</tbody>
</table>