**SUMMARY**

The shared tasks
- News translation, *constrained* (only provided data are used)
- Estonian-English (et-en), Finnish-English (fi-en), Turkish-English (tr-en), and Chinese-English (zh-en)

Key points
- No specific linguistic pre-processing
- SMT and NMT (transformer) systems
- Incremental training for NMT
- Combination of SMT and NMT through n-best lists reranking

The results
- Best systems (BLEU) for Estonian-English and Finnish-English

**INCREMENTAL TRAINING FOR NMT**

Training on more and better back-translated data

Experiments with tr-en
- T-200k-2: r = 2, i₀=200k
- T-100k-1.5: r = 1.5, i₀=100k
- T-fixed: no incremental training
- R-balanced: backtr. with RNN

**DATA**

**Bilingual data**

<table>
<thead>
<tr>
<th>Language pair</th>
<th>#sent. pairs</th>
<th>#tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>et-en</td>
<td>1.9M</td>
<td>29.4M (En)</td>
</tr>
<tr>
<td>fi-en</td>
<td>3.1M</td>
<td>52.9M (Fi)</td>
</tr>
<tr>
<td>tr-en</td>
<td>207.4k</td>
<td>4.6M (Tr)</td>
</tr>
<tr>
<td>zh-en</td>
<td>24.8M</td>
<td>509.9M (Zh)</td>
</tr>
</tbody>
</table>

**Monolingual data**

<table>
<thead>
<tr>
<th>Language</th>
<th>#lines</th>
<th>#tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>338.7M</td>
<td>7.5B</td>
</tr>
<tr>
<td>et</td>
<td>146.1M</td>
<td>3.6B</td>
</tr>
<tr>
<td>fi</td>
<td>177.1M</td>
<td>3.2B</td>
</tr>
<tr>
<td>tr</td>
<td>105.0M</td>
<td>1.8B</td>
</tr>
<tr>
<td>zh</td>
<td>130.5M</td>
<td>2.3B</td>
</tr>
</tbody>
</table>

**Pre-processing**

- Tokenizer:
  - Moses: en, et, fi and tr
  - Jieba: zh
- Truecased (en, et, fi and tr)
- No data filtering

**COMBINATION OF SMT AND NMT**

Method (Marie+, 18)
1. Produce n-best lists with NMT and SMT
2. Compute SMT and NMT features for all the hypotheses
3. Merge the n-best lists
4. Rerank the hypotheses given the feature scores

Features
- Left-right and right-left NMT forward and backward models
- 4-gram language models
- SMT lexical translation probabilities
- Averaged word posterior probabilities
- Hypothesis length features
- System flag
- Phrase-based forced decoding score (for tr-en only) (Zhang+, 18)
- MBR decoding for n-best list (for tr-en only)

**EXPERIMENTS**

**Human evaluation**
- First rank for all tasks (except for zh-en)
- Many other systems are also ranked first
- Strong correlation with BLEU

<table>
<thead>
<tr>
<th>#</th>
<th>System</th>
<th>et→en</th>
<th>en→et</th>
<th>fi→en</th>
<th>en→fi</th>
<th>tr→en</th>
<th>en→tr</th>
<th>zh→en</th>
<th>en→zh</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Moses</td>
<td>18.2</td>
<td>15.1</td>
<td>15.8</td>
<td>10.7</td>
<td>12.1</td>
<td>8.4</td>
<td>16.9</td>
<td>28.0</td>
</tr>
<tr>
<td>2.</td>
<td>Moses NMT-reranked</td>
<td>20.2</td>
<td>17.6</td>
<td>17.5</td>
<td>12.2</td>
<td>14.2</td>
<td>10.1</td>
<td>19.0</td>
<td>29.9</td>
</tr>
<tr>
<td>3.</td>
<td>Marian single (w/o backtr)</td>
<td>22.9</td>
<td>18.5</td>
<td>17.6</td>
<td>13.2</td>
<td>20.2</td>
<td>12.2</td>
<td>23.7</td>
<td>33.0</td>
</tr>
<tr>
<td>4.</td>
<td>Marian single (w/ backtr)</td>
<td>28.6</td>
<td>24.0</td>
<td>23.1</td>
<td>16.8</td>
<td>25.2</td>
<td>18.0</td>
<td>24.7</td>
<td>37.2</td>
</tr>
<tr>
<td>5.</td>
<td>Marian ensemble (w/ backtr)</td>
<td>29.1</td>
<td>24.3</td>
<td>23.6</td>
<td>17.3</td>
<td>25.8</td>
<td>18.3</td>
<td>25.9</td>
<td>37.9</td>
</tr>
<tr>
<td>6.</td>
<td>Moses + Marian</td>
<td>30.7</td>
<td>25.2</td>
<td>24.9</td>
<td>18.2</td>
<td>26.9</td>
<td>19.2</td>
<td>26.7</td>
<td>39.7</td>
</tr>
</tbody>
</table>

- **Moses (#1)** underperforms **Marian (#3)** by several BLEU points
- **Moses NMT-reranked (#2)** is slightly below **Marian (#3)** for fi-en
- Large improvements with back-translation (#4) (> +5BLEU)

- Small improvements with ensemble decoding (6 models) (#5)
- Best results when combining SMT and NMT (#6)
- Best systems (#6) of WMT for et-en and fi-en (BLEU)