French-English and German-English shared translation tasks in both directions

- **n-code**: open source Statistical Machine Translation system
  - Source reordering as pre-processing
  - Translation model based on bilingual n-grams
- **Simple and efficient filtering strategy** of the GigaWord.
- **First use of the SOUL** target language model in SMT

⇒ significant improvements with 10-gram models

### Highlights

**n-code**’s model

- 3-gram Tuple LM and 4-gram target word LM
- Four lexicon models (similar to the phrase table)
- Two lexicalized reordering models (predict orientation of next/previous translation unit)
- Weak distance-based distortion model
- Word-bonus and a tuple-bonus models

### Data Pre-processing

- Better normalization tools provide better BLEU scores
- Specific pre-processing for German as source language
- Cleaning noisy data sets (GigaWord)
  - Discard sentences in other languages
  - Remove repeated sentences, or the ones included in the development sets
  - for the monolingual news data, this can reduce the amount of data by a factor 3 or 4
  - Normalize the character set

### Data Filtering

Filtering the GigaWord Corpus
For each side:

- Train a specific language model on a selection of news texts
- Rank sentences according to their perplexity
- Select sentences above a given threshold

**Two thresholds:**

- The upper quartile ⇒ a subset of 25% (6.7M of sentences)
- The median ⇒ a subset of 50%

### Target n-gram Language Model

**Standard 4-gram Back-off Language Models**

- Total of running words: 2.5G in French and 6.2G in English
- Using all the available data.
- Linear interpolation of several LMs
- Specifically tuned for news text of 2010

**The soul LM**

- A large vocabulary continuous space LM.
- Use a clustering tree to structure the output vocabulary.
- The order n can be raised without a prohibitive increase in complexity.

### Baseline Results (newstest2010)

- Filtering the GigaWord corpus for French-English:

<table>
<thead>
<tr>
<th>System</th>
<th>en2fr</th>
<th>fr2en</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td>All</td>
<td>27.4</td>
<td>56.0</td>
</tr>
<tr>
<td>Upper quartile</td>
<td>27.8</td>
<td>56.3</td>
</tr>
<tr>
<td>Median</td>
<td>28.1</td>
<td>56.0</td>
</tr>
</tbody>
</table>

- German-English, impact of the the POS tagger:

<table>
<thead>
<tr>
<th>System</th>
<th>en2de</th>
<th>de2en</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td>RFTagger</td>
<td>22.8</td>
<td>60.1</td>
</tr>
<tr>
<td>TreeTagger</td>
<td>23.1</td>
<td>59.4</td>
</tr>
</tbody>
</table>

### Results with various LMs

- Linear interpolation of 4 SOUL LMs (different re-sampling)
- Initial shortlist of 5k words
- K-means recursive word clustering based on the continuous representation of words (R), depth of tree = 3
- n-best rescoring, tuned on newstest2009

<table>
<thead>
<tr>
<th>SOUL LM</th>
<th>en2fr</th>
<th>en2de</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td>without</td>
<td>28.1</td>
<td>56.0</td>
</tr>
<tr>
<td>4-gram</td>
<td>28.3</td>
<td>55.5</td>
</tr>
<tr>
<td>6-gram</td>
<td>28.7</td>
<td>55.3</td>
</tr>
<tr>
<td>10-gram</td>
<td>28.8</td>
<td>55.2</td>
</tr>
</tbody>
</table>