We submitted systems for three translation directions: German-to-English, Russian-to-English and English-to-Russian. The focus of our approaches lies on effective usage of the in-domain parallel training data combined with simple scaling of the language and translation models. We use the training data to tune parameter weights for millions of sparse lexicalized features using efficient parallelized stochastic learning techniques. For German-to-English we incorporate syntax features. We combine all systems with large general-domain language models; For RU→EN we use more unfiltered data for the TM.

Sparse, lexicalized features attached to SCFG rules

\[g(x_1) > g(x_2) \iff f(x_1) > f(x_2)\]
\[g() \text{ (per-sentence) BLEU score}\]
\[f() \text{ model score of the decoder}\]
\[w \text{ weight vector}\]
\[x \cdot y \text{ vector dot product}\]

Hinge loss for a stochastic pairwise-ranking perceptron

\[L_w(x) = \max(0, -x \cdot w)\]
\[\nabla L_w(x) = \begin{cases} x, & \text{if } x \cdot w \leq 0, \\ 0, & \text{otherwise.} \end{cases}\]

Gold standard ranking: BLEU+1 scores of translations of kbest lists

Tuning on the training set with \(\ell_1/\ell_2\) regularization and parallelization

(Simianer et al, 2012)

SMT Setup

• cdec SCFG decoder (Dyer et al, 2009)
• Word alignments with a variant of IBM’s model 2 (Dyer et al, 2013)
• Hiero grammars (2 non-terminals max, . . .) built with impl. of the suffix array extraction technique of (Lopez, 2007)
• Language models built with 1mlz (Heatfield, 2013)
• Tokenization, compound splitting and recasing with moses tools

(5) Indicate if spans in decoder derivations match or cross constituents of syntactic trees

In contrast to the syntax feature in Chiang’s original Hiero paper these features do include the actual phrase labels

Effects of soft-syntactic constraints

(2) Derivation using soft-syntactic constraints depicting a sensible parse tree

(Large) Language and Translation Models

German-to-English TM: just TED data \(\Rightarrow\) about 150,000 tokens

English LM: 10\(^{th}\) FR-EN, Europarl, News Commentary, News Crawl, UN corpus, LDC2011T07 \(\Rightarrow\) 7,245,227,502 tokens

Russian-to-English TM: Common Crawl, Yandex 1M, News Commentary, Wiki Headlines, TED data \(\Rightarrow\) 44,042,275 Russian and 48,677,800 English tokens

Russian LM: Common Crawl, News Commentary, Yandex 1M, News Crawl, TED data \(\Rightarrow\) 335,023,785 tokens

Development Results (tst2010)

results on tst2010: \* primary; \* secondary submission; baseline is a standard system with dense features trained with MERT on the dev set

German-to-English:

<table>
<thead>
<tr>
<th>System</th>
<th>TED 4-gram LM</th>
<th>Large 5-gram LM</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>28.7</td>
<td>+1.7</td>
</tr>
<tr>
<td>dtrain-dev</td>
<td>+0.9</td>
<td>+2.1</td>
</tr>
<tr>
<td>dtrain-train (clustered)</td>
<td>+1.3</td>
<td>+2.9</td>
</tr>
<tr>
<td>dtrain-train + soft-syntax</td>
<td>+1.4</td>
<td>-</td>
</tr>
</tbody>
</table>