Hybrid machine translation: Combining rule-based and statistical MT systems

Andreas Eisele
Saarland University & DFKI, LT Lab
Joint work with Kamran Azam, Yu Chen, Christian Federmann, Greg Gulrajani, Eva Hasler, Michael Jellinghaus, Robert Neßelrath, Armin Schmidt, Silke Theison

e-mail: eisele@dfki.de
WWW:http://www.dfki.de/LT
Structure of presentation

- Complementary advantages of rule-based and statistical MT
- Using a SMT decoder to merge outputs of multiple MT engines
- Feeding SMT lexicons into rule-based MT engines
- Thoughts on deeper integration
## EuroMatrix: situation in early 2005

### MT systems per language pair

(Data taken from J. Hutchins’ Compendium of Translation Software, 12th Edition)

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>English</th>
<th>German</th>
<th>French</th>
<th>Spanish</th>
<th>Italian</th>
<th>Portuguese</th>
<th>Dutch</th>
<th>Polish</th>
<th>Lithuanian</th>
<th>Latvian</th>
<th>Greek</th>
<th>Czech</th>
<th>Hungarian</th>
<th>Swedish</th>
<th>Finnish</th>
<th>Slovak</th>
<th>Romanian</th>
<th>Danish</th>
<th>Bulgarian</th>
<th>Slovene</th>
<th>Maltese</th>
<th>Lithuanian</th>
<th>Irish</th>
<th>Estonian</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>47</td>
<td>41</td>
<td>44</td>
<td>30</td>
<td>30</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>German</td>
<td>48</td>
<td>24</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>French</td>
<td>40</td>
<td>23</td>
<td>11</td>
<td>13</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Spanish</td>
<td>41</td>
<td>7</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Italian</td>
<td>29</td>
<td>10</td>
<td>13</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Portuguese</td>
<td>29</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dutch</td>
<td>10</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Polish</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Latvian</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Greek</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Czech</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Hungarian</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Swedish</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Finnish</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Slovak</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Romanian</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Danish</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bulgarian</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Slovene</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Maltese</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Lithuanian</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Irish</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Estonian</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Most language pairs remain uncovered
### EuroMatrix: situation in early 2005

MT systems per language pair (data taken from J.Hutchins’ Compendium of Translation Software, 12th Edition)

<table>
<thead>
<tr>
<th></th>
<th>Eng</th>
<th>Germ</th>
<th>Fre</th>
<th>Spa</th>
<th>Port</th>
<th>Pol</th>
<th>Lat</th>
<th>Gre</th>
<th>Cze</th>
<th>Hung</th>
<th>Swe</th>
<th>Fin</th>
<th>Slo</th>
<th>Rom</th>
<th>Dan</th>
<th>Bel</th>
<th>Slo</th>
<th>Mal</th>
<th>Lith</th>
<th>Irish</th>
<th>Est</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>![Flag]</td>
<td>47</td>
<td>41</td>
<td>44</td>
<td>30</td>
<td>30</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>48</td>
<td>![Flag]</td>
<td>24</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>40</td>
<td>-</td>
<td>![Flag]</td>
<td>8</td>
<td>10</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>41</td>
<td>Amikai</td>
<td>Babelfish</td>
<td>Click2Translate</td>
<td>Dictionary.com</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>29</td>
<td>Translator</td>
<td>Easy Translator</td>
<td>e-Translation Server</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td>29</td>
<td>FB-Active</td>
<td>FB-Win</td>
<td>FJWSpilltrans</td>
<td>FreeTranslation</td>
<td>GETrans</td>
<td>Google</td>
<td>Hypertrans</td>
<td>IM Translator</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>10</td>
<td>iTranslator On-line</td>
<td>JxEuro</td>
<td>Korya</td>
<td>Ewa Ippatsu</td>
<td>Honyaku</td>
<td>Language Weaver SMTS</td>
<td>LocalTranslation</td>
<td>LogoMedia</td>
<td>Lycos</td>
<td>MZ-Win Translator</td>
<td>NeuroTran</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>2</td>
<td>Global</td>
<td>Pragma</td>
<td>Pragma Online</td>
<td>@promt</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>2</td>
<td>PROMT-Online</td>
<td>PT-SMS</td>
<td>PT-WAP</td>
<td>Reverso [series]</td>
<td>SDL Enterprise</td>
<td>Smart Translator</td>
<td>Systran</td>
<td>T1</td>
<td>Transcend</td>
<td>translate</td>
<td>Translation</td>
<td>TranSphere</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovak</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>1</td>
<td>Web-Transer BB Multilingual</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Danish</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bulgarian</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovene</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maltese</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lithuanian</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Irish</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estonian</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Most language pairs remain uncovered, but some are densely populated.
Rule-based vs. statistical MT

Some examples (translate pro ↔ SMT Koehn 2005)

EN: I wish the negotiators continued success with their work in this important area.

DE: Ich wünsche den Unterhändlern auch weiterhin viel Erfolg auf diesem wichtigen Gebiet.

RBMT: Ich wünsche, dass die Unterhändler Erfolg mit ihrer Arbeit in diesem wichtigen Bereich fortsetzten.

SMT: Ich wünsche der Verhandlungsführer fortgesetzte Erfolg bei ihrer Arbeit in diesem wichtigen Bereich.

continued: verb instead of adjective

three inflectional endings wrong
More examples

EN: We seem sometimes to have lost sight of this fact.

RBMT: Wir scheinen manchmal Anblick dieser Tatsache verloren zu haben.

idiomatic expression not known

SMT: Manchmal scheinen wir aus den Augen verloren haben, diese Tatsache.

wrong attachment of „diese Tatsache“
Rule-based vs. statistical MT

More examples

EN: I would like to close with a procedural motion.

DE: Mit einem Antrag zur Geschäftsordnung komme ich zum Schluss.

RBMT: Ich möchte mit einer verfahrenstechnischen Bewegung schließen.

wrong translation of procedural motion

SMT: Ich möchte abschließend eine Frage zur Geschäftsordnung.

verb is missing
Rule-based vs. statistical MT

More examples

EN: The leaders of Europe have not formulated a clear vision.

DE: Die Führung Europas hat keine klaren Visionen formuliert.

RBMT: Die Leiter von Europa haben keine klare Vision formuliert.

„Leiter von Europa“ sounds very odd

SMT: Die Führung Europas nicht formuliert eine klare Vision.

syntactically illformed
In the early 90s, statistical and rule-based approaches were seen in strict contrast. But PROs and CONs are complementary:

<table>
<thead>
<tr>
<th>Method</th>
<th>Syntax</th>
<th>Structural Semantics</th>
<th>Lexical Semantics</th>
<th>Lexical Adaptivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based MT</td>
<td>++</td>
<td>+</td>
<td>-</td>
<td>--</td>
</tr>
<tr>
<td>Statistical MT</td>
<td>--</td>
<td>--</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Example-based MT</td>
<td>-</td>
<td>--</td>
<td>-</td>
<td>++</td>
</tr>
</tbody>
</table>

⇒ It is now more or less consensus to target integrated approaches.
Deep Integration: Design a new setup that combines the advantages of two paradigms, e.g. by integrating some good features of Approach B into Approach A, such as

- Making a rule-based system adaptive by adding a module for rule learning
- Making a SMT system syntax-aware by adding syntactical constraints/rules

Shallow Integration: Integrate two or more systems, following different approaches, into a larger system

Deep integration is superior in the long run, but also much more challenging. WP2 (Richer models for statistical translation, U Edinburgh) and WP3 (Tree-Based Transfer Models, Charles U Prague) are steps towards deep integration.

WP6 (Saarland U) will pursue the modest goal of shallow integration into a MEMT architecture, hopefully still giving insights into the relevant issues related to deep integration.
Multi-engine MT via black-box integration
(as done in VerbMobil and earlier)

Source Text → Rule-based MT engines → Hypotheses → Selection → Target Text

SMT-engine(s)
Towards better hybrid MT Architectures

- Disadvantage of simple selection:
  For longer sentences, no result will be perfect; we want to combine better parts of multiple outcomes.

- But recombination can be fairly complex, as corresponding parts of alternative candidates are not obvious

- We need
  - alignment of MT results (needs to cope with MT errors)
  - search for best combination

- We can use existing SW modules for both purposes in first steps, error analysis may then suggest improvements
Hybrid MT Architecture II

- Parallel Corpus
- Alignment, Phrase Extraction
- Monolingual Corpus
- Counting, Smoothing

SMT has the last word

- Phrase-Table
- nGram-Model
- Dyn. PT
- SMT Decoder

- Target Text
- Source Text

- Rule-based MT engines
- Hypotheses

1st MT Marathon, Edinburgh, April 17, 2007
Hybrid MT Architecture II

Key features:
- Source text is sent through many MT engines, including web-based and locally installed ones.
- Alignment between MT output and source text is done via (modified) GIZA++.
- Alignment quality is improved by using models trained on larger datasets.
- Alignment waiting times are reduced by client-server setup.
Hybrid MT Architecture II

Current status:

- Using 6..7 engines: Systran, SDL, ProMT, OpenLogos, translate pro, L&H PTP, Lucy
- Each of the MT engines has its own peculiarities that require attention (encoding, tokenisation, …)
- Combined phrasetables slow down decoding, makes MERT more difficult
- Delays due to technical problems while preparing WMT07 submission
- Implementation essentially done, but current configuration does not yet beat baseline
Hybrid MT Architecture II

Indicative results

PoS-aware BLEU-1 score

<table>
<thead>
<tr>
<th>Systems</th>
<th>Overall (%)</th>
<th>Named Entities (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-I</td>
<td>51.36</td>
<td>34.52</td>
</tr>
<tr>
<td>R-II</td>
<td>51.34</td>
<td>64.84</td>
</tr>
<tr>
<td>SMT</td>
<td>55.55</td>
<td>20.90</td>
</tr>
<tr>
<td>Hybrid</td>
<td>55.53</td>
<td>49.53</td>
</tr>
</tbody>
</table>

Ratio of untranslated tokens

<table>
<thead>
<tr>
<th>Systems</th>
<th>Token #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref.</td>
<td>2091 (4.21%)</td>
</tr>
<tr>
<td>R-I</td>
<td>3886 (7.02%)</td>
</tr>
<tr>
<td>R-II</td>
<td>3508 (6.30%)</td>
</tr>
<tr>
<td>SMT</td>
<td>3976 (7.91%)</td>
</tr>
<tr>
<td>Hybrid</td>
<td>2425 (5.59%)</td>
</tr>
</tbody>
</table>
Hybrid MT Architecture III

Rule-based MT engine

Source Text

Target Text

MT Lexicon

Alignment, Phrase Extraction

Phrase-Table

Parallel Corpus

Linguistic Processing,

SMT feeds rule-based MT
Hybrid MT Architecture III

Key issues:

- RBMT has no mechanism to disprefer implausible results → lexicon needs to be correct
- MT lexicon needs information not contained in parallel texts (lemma, gender, inflection, …)
- Current tools make fully automatic high-quality lexicon extraction rather difficult

Architecture requires manual effort → make it as simple and generic as possible, e.g. by using encoding standard for lexical data like OLIF
Hybrid MT Architecture III

Source Text → Rule-based MT engine → Target Text

MT Lexicon

Manual Validation

Alignment, Phrase Extraction

Phrase-Table

Parallel Corpus

SMT feeds rule-based MT

Linguistic Processing,
OLIF standard has been developed to facilitate exchange of multilingual lexical data.

- Contains encodings for part-of-speech and head, morphosyntactic features, and inflectional behaviour
- Current version 2.1 has focus on English, French, German, Portuguese, and Spanish
- See [www.olif.net](http://www.olif.net) for details
Ongoing cooperation between DFKI and European Patent Office (EPO)

- Goal: Extract parallel terminologies for EN, DE, ES, FR from translated patent documents
- Motivation for EPO: Build up infrastructure for machine translation of patents, acquire relevant technical vocabulary
- Motivation for DFKI: Develop industrial applications of techniques from MT research, strengthen NLP tools
Technical documentation makes up a large share of language industry’s raw material, vocabulary is commercially interesting.

Manual construction of broad-coverage or unrestricted multilingual terminologies would be prohibitively expensive.

Translated documents exist in large volumes, as well as techniques for sentence/word/phrase alignment, these can be used to partially automate the task.

IPC (hierarchical system of about 70K classes) may help to relate extracted terms with ontologies.

Test-bed for scalability of tools and resources:
- How well do our tools cover technical texts?
- Can we acquire new lexical information from data?

First step towards MT for technical documents.
Terminology Extraction from Patents

History and current status:

- Techniques were prototypically implemented in a feasibility study for WIPO ('03, via acrolinx GmbH)
- Call for Tender by EPO in August ’05, bids and results on test data due in September
- From 14 bids, DFKI delivered best results for DE↔EN, ES↔EN and among the best for FR↔EN
- Test phase December ’05..July ’06: Term extraction from samples, feasibility study on validation
- Production phase (since August ’06): Work on 50 million sentence pairs (~ 2E9 running words), manual validation of specific subsets
- Continuation in 2007 may broaden scope to additional languages: PT, IT, RO, NL, SW
The International Patent Classification (IPC)

- based on the Strasbourg Agreement (1971)
  used by >100 national authorities
- indispensable for finding prior art
- hierarchical structure, consisting of
  - eight sections (A..H)
  - 120 classes (A01 … H05)
  - 628 subclasses (A01B…H05K)
  - ≈69,000 subdivisions (e.g. A01B 1/02 or H05K 10/00)
- regularly updated (currently in force: 8th edition)
- officially released in EN and FR by WIPO, but translations to many languages are available from national authorities
The International Patent Classification (IPC)

- based on the Strasbourg Agreement (1971)
- used by >100 national authorities
- indispensable for finding prior art
- hierarchical structure, consisting of:
  - eight sections (A..H)
  - 120 classes (A01 … H05)
  - 628 subclasses (A01B…H05K)
  - ≈69,000 subdivisions (e.g. A01B 1/02 or H05K 10/00)
- regularly updated (currently in force: 8th edition)
- officially released in EN and FR by WIPO, but translations to many languages are available from national authorities
The International Patent Classification (IPC) is indispensable for finding prior art. It is based on the Strasbourg Agreement (1971), is used by >100 national authorities, and has a hierarchic structure consisting of:

- eight sections (A..H)
- 120 classes (A01 … H05)
- 628 subclasses (A01B…H05K)
- ≈69,000 subdivisions (e.g. A01B 1/02 or H05K 10/00)

It is regularly updated (currently in force: 8th edition) and officially released in EN and FR by WIPO, but translations to many languages are available from national authorities.
The International Patent Classification (IPC)

- based on the Strasbourg Agreement (1971)
- used by >100 national authorities
- indispensable for finding prior art
- hierarchical structure
  - eight sections (A..H)
  - 120 classes (A01 … H05)
  - 628 subclasses (A01B…H05K)
  - ≈69,000 subdivisions (e.g. A01B 1/02 or H05K 10/00)
- regularly updated (currently in force: 8th edition)
- officially released in EN and FR by WIPO, but translations to many languages are available from national authorities
Terminology Extraction from Patents

Research questions related to the IPC

■ Automatic Classification
   Can IPC classes be identified automatically?
   *(So far classification and search done by ~ 6500 experts)*

■ Ontology construction
   How does the IPC relate to the terminologies used in the various domains? Can we (semi-) automatically construct/extend these terminologies given the documents?

■ Word sense disambiguation
   Can a given IPC class help to identify meaning/translation of a given term?
Technical setup:
- Use linguistic tools for corpus annotation
  - POS-tagging, phrase recognition, lemmatization
- Use statistical tools for alignment
  - GIZA++ from Franz Och
  - Own algorithms based on word similarities
- Integrate module outcomes, transform into OLIF entries

Improvement in 2\textsuperscript{nd} phase:
- Feed-back of modifications to basic modules
- Infrastructure for manual validation
- Manual inspection and error analysis is used to improve algorithms as long as the project is ongoing
Terminology Extraction: Architecture

- Statistical Word Alignment
- Word-Level Matches
- Integration
- Phrase-Level Matches
- Selection and OLIF transformation
- Augmented Documents (POS, chunks, lemmata)
- OLIF DB
- Parallel Documents
- Linguistic Processing
Examples for Patent Terminology

Postbestimmungsortinformationsspeichereinrichtung
  = mail destination information memory means
Informationsdurchforstungssteuerungseinrichtung
  = information browsing control means
Hypervideonachrichtversendungsverarbeitungseinrichtung
  = hypervideo message posting processing means
Gasphasenverunreinigungsabsorptionsflüssigkeit
  = gas phase contaminant absorbing liquid
Manual Validation of Terminology

- Original Plan:
  - Validation by (30..40) domain experts in national patent offices, but:
  - Linguistic validation not suitable for patent examiners
- New setup: Validation work is shared between
  - DFKI for linguistic validation and
  - patent offices for domain knowledge
- Validation workflow handled in a Web-based infrastructure for terminology maintenance
  - Prototype available since Fall ’06
  - Successfully used for first deliveries
Manual Validation of Terminology
Next steps:

- Use existing infrastructure to feed various rule-based MT engines (OpenLogos, Lucy)
- Measure impact on results
- Decide on domain for which extended lexicons would be most useful
Conclusion

- We have presented two complementary architectures to combine rule-based and statistical MT engines.
- Implementation is fairly advanced but fine-tuning still needs to be done.
- These setups can themselves be combined into a MEMT system.
- Truly deep integration using rule-based and statistical knowledge sources in well-balanced way will need more work.
Thank You for Your Attention