MT Marathon

26th – 30th January 2009

Prague, Czech Republic
## Lectures, Talks, Labs

<table>
<thead>
<tr>
<th>Time</th>
<th>Morning Lecture</th>
<th>Research Talks</th>
<th>Afternoon Lab Session</th>
<th>Evening</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.00 – 12.30</td>
<td>Word Alignment (Barry Haddow)</td>
<td>PostCAT, apertium-cy, MBMT</td>
<td>Implementing IBM model 1</td>
<td>Projects: Update</td>
</tr>
<tr>
<td>14.00 – 17.00</td>
<td>Phrase-Based Models and Decoding (Chris Dyer)</td>
<td>Joshua, SAMT, MERT+</td>
<td>Installing and running Moses (Hieu Hoang and Josh Schroeder)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>TectoMT: Processing Trees (Zdeněk Žabokrtský and others)</td>
<td>RIA, Sub-Tree Aligner</td>
<td>TectoMT hands-on experience: Installation and tutorial (Jana Kravalová)</td>
<td>Projects: Update</td>
</tr>
<tr>
<td></td>
<td>Richer Models and Optimization (Philipp Koehn)</td>
<td>Projects: Final Short Presentations</td>
<td>Using factored models and MERT in Moses (Hieu Hoang, Barry Haddow, Abhishek Arun)</td>
<td></td>
</tr>
</tbody>
</table>

### Accepted Contributions, Research Talks

The following contributions will be presented during late mornings:

1. **apertium-cy**: F. M. Tyers, K. Donnelly: apertium-cy - a collaboratively-developed free RBMT system for Welsh to English
3. **MBMT**: A. van den Bosch and P. Berck: Memory-Based Machine Translation and Language Modeling
4. **MERT+**: N. Bertoldi, B. Haddow, J.-B. Fouet: Improved Minimum Error Rate Training in Moses
5. **PostCAT**: J. Graça, K. Ganchev, B. Taskar: PostCAT - Posterior Constrained Alignment Toolkit
7. **SAMT**: A. Venugopal, A. Zollmann: Grammar based statistical MT on Hadoop. An end-to-end toolkit for large scale PSCFG based statistical machine translation
8. **Sub-Tree Aligner**: V. Zhechev: Unsupervised Generation of Parallel Treebanks through Sub-Tree Alignment
9. **Z-MERT**: O. Zaidan: Z-MERT: A Fully Configurable Open Source Tool for Minimum Error Rate Training of Machine Translation Systems (There is no presentation for this paper.)

Research talk presentations should be 20 to 25 minutes long with additional 5 minutes for a discussion.
Contents

Monday
Introduction to MT ......................................................... 4
MT Evaluation ............................................................... 13
Intro to Stat-XFER .......................................................... 19
Stat-XFER: Czech to English Translation Project ......................... 29
TectoMT for Plaintext Freaks ............................................. 33

Tuesday 35
Word Alignment ................................................................. 35

Wednesday 42
Phrase-Based Models and Decoding ...................................... 42

Thursday 51
Introduction to the Software Framework ................................ 51
TectoMT: Alignment .......................................................... 57
Bad news, NLP Hacking and Feature Fishing ............................ 59
TectoMT: Tutorial .............................................................. 62

Friday 69
Discriminative Training and Factored Translation Models ............. 69