Surprise Language Evaluation: 
Rapid-Response Cross-Language IR

**Maryland:** Douglas W. Oard, Marine Carpuat, Petra Galuscakova, 
Joseph Barrow, Suraj Nair, Xing Niu, Han-Chin Shing, 
Weijia Xu, Elena Zotkina

**Columbia:** Kathleen McKeown, Smaranda Muresan, Efsun Selin Kayi, 
Ramy Eskander, Chris Kedzie, Yan Virin

**Yale:** Dragomir Radev, Rui Zhang

**Cambridge:** Mark Gales, Anton Ragni

**Edinburgh:** Kenneth Heafield

June 9, 2019   EVIA 2019
Looking Backward

• 1966: ALPAC report
  – Refocus investment on enabling technologies

• 1988: IBM’s Candide MT system
  – Data-driven approach

• 2003: DARPA TIDES surprise languages
  – Cebuano and Hindi

• 2017: IARPA MATERIAL program
Surprise Language Exercises

<table>
<thead>
<tr>
<th></th>
<th><strong>TIDES (2003)</strong></th>
<th><strong>MATERIAL (2019)</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>English users / Docs in X</td>
<td>Time constrained</td>
<td>English users / Docs in X</td>
</tr>
<tr>
<td></td>
<td>Research-oriented</td>
<td>Time constrained</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Research-oriented</td>
</tr>
<tr>
<td></td>
<td>Zero-resource start</td>
<td>Language pack start</td>
</tr>
<tr>
<td></td>
<td>Digital text</td>
<td>Digital text and speech</td>
</tr>
<tr>
<td></td>
<td>Collaborative</td>
<td>Competitive</td>
</tr>
</tbody>
</table>
# TIDES Schedule (2003)

<table>
<thead>
<tr>
<th>Cebuano</th>
<th>Hindi</th>
</tr>
</thead>
<tbody>
<tr>
<td>Announce: March 5</td>
<td>Jun 1</td>
</tr>
<tr>
<td>Test Data:</td>
<td>Jun 27</td>
</tr>
<tr>
<td>Stop Work: March 14</td>
<td>Jun 30</td>
</tr>
</tbody>
</table>
Cebuano Resources

- Bible: 913K words
- Examples of usage: 214K words (OCR)
- Communist Party Newsletter: 138K words
- Term list: 20K entries
- Web pages: 58K words
- Manual news translation
  - Discriminative training: 6K words
  - MT Evaluation: 13K words
Example Cebuano Translation

question transparent is our government?
of salem arellano, mindanao scoop, 17 november 2002
of so that day the seminar that was held in america
that from the four big official of the seven the place in
mindanao run until is in davao. the purpose of the
seminar, added of members orlando maglinao, is the
resistance to cause the corruption in the government
is be, ue, ue of our country.
Cebuano CLIR at Maryland

• Starting Point: iCLEF 2002 German system
  – Interface: “synonyms”/examples (parallel)/MT
  – Back end: InQuery/Pirkola’s method

• 3-day porting effort
  – Cebuano indexing (no stemming)
  – One-best gloss translation (bilingual term list)

• Informal Evaluation
  – 2 Cebuano native speakers (at ISI)
Hindi Results

• Several components
  o POS tags, morphology, time expression, parsing

• 5 evaluated tasks
  o CLIR (English queries)
  o Topic tracking (English examples)
  o Machine translation into English
  o English “Headline” generation
  o Entity tagging

• 5 demos
  o Interactive CLIR (2 systems)
  o Cross-language QA
  o Machine translation
  o Cross-document entity tracking
Hindi Resources

• Much more content available than for Cebuano
  – Total: 4.2 million words

• Large and diverse
  – Web, news, dictionaries, handbooks, hand translated, …

• Huge effort: data conversion/cleaning/debugging
  – Many non-standard encodings
  – Often: no converters available or available converters do not work properly
Translation Elicitation Server  
- Johns Hopkins University (David Yarowsky)

- People voluntarily translated large numbers of Hindi news sentences for nightly prizes at a novel Johns Hopkins University website
- Performance is measured by Bleu score on 20% randomly interspersed test sentences
- Allows immediate way to rank and reward quality translations and exclude junk
- Result: 300,000 words of perfectly sentence-aligned bitext (exactly on genre) for 1-2 cents/word within ~5 days
  - Much cheaper than 25 cents/word for translation services or 5 cents/word for a prior MT-group’s recruitment of local students

Sample Interface:

User choice of 2-3 encoding alternatives

- Observed exponential growth in usage (before prizes ended)
  - viral advertising via family, friends, newgroups, ...
- $0 in recruitment, advertising, and administrative costs
  - Nightly incentive rewards given automatically via amazon.com gift certificates to email addresses (any $ amount, no fee)
  - no need for hiring overhead. Rewards only given for proven high quality work already performed (prizes not salary).
  - immediate positive feedback encourages continued use
- Direct immediate access to worldwide labor market fluent in source language
Example Hindi Translation

• Indonesian City of Bali in October last year in the bomb blast in the case of imam accused India of the sea on Monday began to be averted. The attack on getting and its plan to make the charges and decide if it were found guilty, he death sentence of May.
Indonesia of the police said that the imam sea bomb blasts in his hand claim to be accepted. A night Club and time in the bomb blast in more than 200 people were killed and several injured were in which most foreign nationals. …
Hindi CLIR

- N-grams (trigrams best for UTF-8)
- Relative Average Term Frequency (Kwok)
- Scanned bilingual dictionary (Oxford)
- More topics for test collection (29)
- Weighted structured queries (IBM lexicon)
- Alternative stemmers (U Mass, Berkeley)
- Blind relevance feedback
- Transliteration
- Noun phrase translation
- MIRACLE integration (ISI MT, BBN headlines)
Hindi CLIR Formative Evaluation

![Graph showing the Mean Reciprocal Rank over days with 19 known item queries.](graph.png)
Some Challenges in 2003

• Formative evaluation

• Synchronize variable-rate efforts
  – More like soccer than football

• Integration

• Capturing lessons learned
MATERIAL in 2019: CLIR Pipeline
## Lithuanian ASR (Cambridge)

<table>
<thead>
<tr>
<th>Day</th>
<th>Description</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>CTS</td>
</tr>
<tr>
<td>1</td>
<td>BABEL OP2 build</td>
<td>48.2</td>
</tr>
<tr>
<td>2</td>
<td>Baseline GMM-HMM</td>
<td>55.2</td>
</tr>
<tr>
<td>3</td>
<td>Baseline NN-HMM</td>
<td>41.1</td>
</tr>
<tr>
<td>4</td>
<td>Web language model</td>
<td>39.1</td>
</tr>
<tr>
<td>5</td>
<td>Speed perturbation</td>
<td>37.9</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>N</td>
<td>More text and audio</td>
<td>35.4</td>
</tr>
</tbody>
</table>

- Systems distributed to the team within 5 days marked in blue
### Lithuanian MT

<table>
<thead>
<tr>
<th>corpus</th>
<th>sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATERIAL-BUILD (train)</td>
<td>42,000</td>
</tr>
<tr>
<td>ParaCrawl-BiCleaner-v4 (train)</td>
<td>815,713</td>
</tr>
<tr>
<td>MATERIAL-BUILD (dev)</td>
<td>1,654</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Edinburgh</th>
<th>newstest2019</th>
<th>BLEU-4</th>
<th>1-gram prec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-lt</td>
<td>21.52</td>
<td>51.4</td>
<td></td>
</tr>
<tr>
<td>lt-en</td>
<td>28.07</td>
<td>60.5</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Maryland</th>
<th>newstest2019</th>
<th>System</th>
<th>BLEU-4</th>
<th>1-gram prec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>en-lt</td>
<td></td>
<td>SMT</td>
<td>13.00</td>
<td>44.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NMT</td>
<td>4.69</td>
<td>22.0</td>
</tr>
<tr>
<td>lt-en</td>
<td></td>
<td>SMT</td>
<td>20.73</td>
<td>56.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NMT</td>
<td>16.25</td>
<td>50.3</td>
</tr>
</tbody>
</table>
Lithuanian 5 Day Exercise: Week 1

Build pack released
- 4:14 PM
- Lithuanian assigned

DBQT CLIR system
- 1:13 PM
- Uses Wiktory
- 0.04 MAP

CLIR test collection
- 5:41 PM

7PSQ CLIR systems
- 8:30 PM
- 0.08 MAP

Neural MT translations
- 4:04 PM

Baseline CLIR system
- 11:35 PM
- No translations used

ASR system
- 5:22 PM
- 55% WER

ASR system
- 5:00 PM
- 41% WER

176 PSQ and DBQT systems
- 6:45 PM
- 0.41 MAP for PSQ SMT

Document translation CLIR
- 8:40 PM
- Uses NMT
- 0.43 MAP
Lithuanian 6 Day Exercise: Week 2

**4/8/2019**
- SMT and another NMT translations 5:50 AM

**4/9/2019**
- System Combination CLIR 10:48 PM
- 0.82 MAP
- 36% WER
- ASR system 4:07 PM
- 10:50 AM
- AQWV=0.44 on Dev for SMT

**4/10/2019**
- 77 document translation systems
  - 11:55 AM
  - AQWV=0.34 for manual translations

**4/11/2019**
- Dev and Analysis released 5:13 PM
- SMT PSQ CLIR system 2:50 PM
  - AQWV=0.21 on Dev
- 52PSQ and DBQT CLIR systems
  - 4:43 PM
  - AQWV=0.40 on Dev for PSQ SMT
- CLIR system combination
  - 4:02 PM
  - MQWV=0.54 on Dev
  - AQWV=0.47 on Analysis

**4/12/2019**
- 10:39 PM
- AQWV=0.19 on Analysis for NMT
Example Summary: “food shortage”

CLOSE MATCH (food shortage):
...can cause stress, some food products, air change, lack of food, misunderstandings, as well as many other factors. Klasterinis headache pain It is quite rare, strong headache, which is more widespread between men than women. Klasterinis headache may arise one time during the day...
Example Summary: “food shortage”

**Machine Translation Summary**

CLOSE MATCH (food shortage):
...can cause stress, some food products, air change, lack of food, misunderstandings, as well as many other factors. Klasterinis headache pain It is quite rare, strong headache, which is more widespread between men than women. Klasterinis headache may arise one time during the day...

**Human Translation Summary**

CLOSE MATCH (food shortage):
...increases; it's often hereditary (to family members in one or several generations); a migraine can be caused by stress, some food products, changes in weather, lack of food, insomnia, also numerous other factors. Cluster headache it's a quite rare, severe headache that's more prevalent among...
Human Evaluation on Query 1 - Analysis

<table>
<thead>
<tr>
<th>CLIR (machine translation)</th>
<th>AQWV</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ E2E (manual translation)</td>
<td>0.34</td>
</tr>
<tr>
<td>+ E2E (machine translation)</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Some Lessons Learned

• Build on:
  – Existing infrastructure
  – Existing team

• Language packs enable rapid progress
  – Reuse it when the core technology improves

• Provide IR eval data on day 1
Lithuanian
Surprise Language
Hall of Fame