Project overview

- QuEst is a framework for quality estimation
  - extracts features from input/output samples
  - trains and applies ML models using those features
  - predicts expected PE effort/editing time/translation quality

- MTM Projects with QuEst
  1. Discourse-level features for quality estimation
  2. Detecting/modelling human translation errors
Participants

1. Lucia Specia
2. Eleftherios Avramidis
3. Carolina Scarton
4. José Guilherme Camargo de Souza
5. Matteo Negri
6. Jie Jiang
7. Mark Fishel
8. Niko Papula
9. Konstantinos Chatzitheodorou
10. Shadi Saleh
11. Christian Buck
First steps

- Participants got acquainted with QuEst and with adding features to the code
- Brainstorming on possible new features for sub-projects (1) and (2)
- Sources of inspiration for features
  - MQM from QTLaunchPad
  - Pre-defined ideas by Carolina Scarton

https://docs.google.com/document/d/1dJNCzoHzCKXEE-c480dRoVugO6RtvCoPrpayvbENYI/edit#heading=h.gjdgxs
Machine Translation Error Corpora

9::source-txt: the commission considers that the submissions do not bring new elements to its assessment.

9::ref-err-cats: komisjon leiab et ettepanekud miss ei sisalda lex uusi elemente lex tema lex hinnangu infl jaoks miss

9::hyp-err-cats: komisjon leiab et ei looda lex uusi andmeid lex oma lex hinnang infl

http://terra.cl.uzh.ch
### Human Translation Error Corpora

<table>
<thead>
<tr>
<th>Original context</th>
<th>en</th>
<th>The members of the family who are admitted into another Member State (under the title) of a &quot;family group&quot; are entitled to equal treatment, i.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternative student context</td>
<td>en</td>
<td>He may not be allowed to take part in the management of bodies governed by public law or to work in a body governed by public law. He is also entitled to eligibility in workers' representation bodies of the company. The family members of any EU (worker's) husband, wife, descendants aged under 21 years old, children and dependents, regardless of their nationality, have the right to settle with the worker in the host Member State.</td>
</tr>
</tbody>
</table>

---

[http://corpus.leeds.ac.uk/mellange/ltc.html](http://corpus.leeds.ac.uk/mellange/ltc.html)
Urdu-English turker translations

- Prepared and annotated with Meteor scores
  a corpus of HT and PE by turkers
  - 1792 Urdu source sentences
  - 4 reference translations by professional translators
  - 4 human translations by turkers
  - 3 human post-editions for each of the 4 turk translations, also by turkers

Zaidan & Callison-Burch, 2011: Crowdsourcing Translation: Professional Quality from Non-Professionals
http://cs.jhu.edu/~ozaidan/RCLMT/UrEn-NIST09_mturk.zip
## Urdu-English turker translations

- Baseline features, predicting Meteor [0, 1]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline - mean of training</th>
<th>Baseline 17 features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turk Translator 1</td>
<td>0.1790</td>
<td>0.1669</td>
</tr>
<tr>
<td>Turk Translator 2</td>
<td>0.1986</td>
<td>0.1771</td>
</tr>
<tr>
<td>Turk Translator 3</td>
<td>0.2111</td>
<td>0.2084</td>
</tr>
<tr>
<td>Turk Translator 4</td>
<td>0.1756</td>
<td>0.1700</td>
</tr>
<tr>
<td>Turk Editor 1-1</td>
<td>0.1800</td>
<td>0.1692</td>
</tr>
<tr>
<td>Turk Editor 1-2</td>
<td>0.1844</td>
<td>0.1727</td>
</tr>
<tr>
<td>Turk Editor 1-3</td>
<td>0.1910</td>
<td>0.1867</td>
</tr>
</tbody>
</table>
**Urdu-English turker translations**

- Baseline features, predicting Meteor \([0,1]\)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Baseline - mean of training</th>
<th>Baseline 17 features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Turk Editor 2-1</td>
<td>0.1854</td>
<td>0.1760</td>
</tr>
<tr>
<td>Turk Editor 2-2</td>
<td>0.2005</td>
<td>0.1824</td>
</tr>
<tr>
<td>Turk Editor 2-3</td>
<td>0.2038</td>
<td>0.1876</td>
</tr>
<tr>
<td>Turk Editor 3-1</td>
<td>0.2027</td>
<td>0.1983</td>
</tr>
<tr>
<td>Turk Editor 3-2</td>
<td>0.2126</td>
<td>0.2077</td>
</tr>
<tr>
<td>Turk Editor 3-3</td>
<td>0.2103</td>
<td>0.2014</td>
</tr>
<tr>
<td>Turk Editor 4-1</td>
<td>0.1700</td>
<td>0.1660</td>
</tr>
</tbody>
</table>
New features for human transl. QE

- **Jie Jiang:**
  - # and % slang words in target language
  - # and % of abbreviation conflicts in target language

- **Shadi Saleh and Niko Papula:**
  - # of NE in source text (using Stanford NER at feature extraction time)

- **Mark Fishel:**
  - # phrases from a monolingual terminology
  - # correctly and incorrectly translated phrases acc. to a bilingual terminology

- **Jose G.C. de Souza and Matteo Negri:**
  - Approximation of terminology similarity using cross-lingual entropy techniques (based on LMs)
New features for discourse-wide QE

● Carolina Scarton and Lucia Specia:
  ○ Token repetition in target
  ○ Lemma repetition in target
  ○ Token repetition in source
  ○ Lemma repetition in source
  ○ Ratio of token repetition (target/source)
  ○ Ratio of lemma repetition (target/source)
  ○ Noun repetition in target
  ○ Noun repetition in source
  ○ Ratio of noun repetition (target/source)
New features for discourse-wide QE

- Corpus of **subtitles** (en-pt) with episodes as documents (6, 2 per series)
- Predicting HTER scores [0,1]
- Results for 1 (out of 4) MT system:

<table>
<thead>
<tr>
<th>MAE</th>
<th>Baseline - mean of training</th>
<th>Baseline 17 features</th>
<th>17 + Discourse features</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3389</td>
<td>0.2324</td>
<td>0.2219</td>
<td></td>
</tr>
</tbody>
</table>
New data

- Human translation QE
  - Turkers corpus (Lucia Specia)
  - MeLLANGE corpus with word-level tagging (Mark Fishel, Christian Buck)
    - 232 texts of ~350 words each
    - Annotation at word level
    - Content and language errors

- Discourse-wide QE
  - Subtitles corpus (Carolina Scarton, Lucia Specia)
New resources

- Slang list (Jie Jiang)
- Abbreviation list (Jie Jiang)
- EuroVoc “terminology” database (Konstantinos Chatzitheodorou)
- Style guide for EU languages with respect to dates, numbers, etc. (Konstantinos Chatzitheodorou)
- Currency codes (Konstantinos Chatzitheodorou)
Code

- New branch on github with some code already merged (Eleftherios Avramidis)