QuEst @MTM

Report
September 14, 2013

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

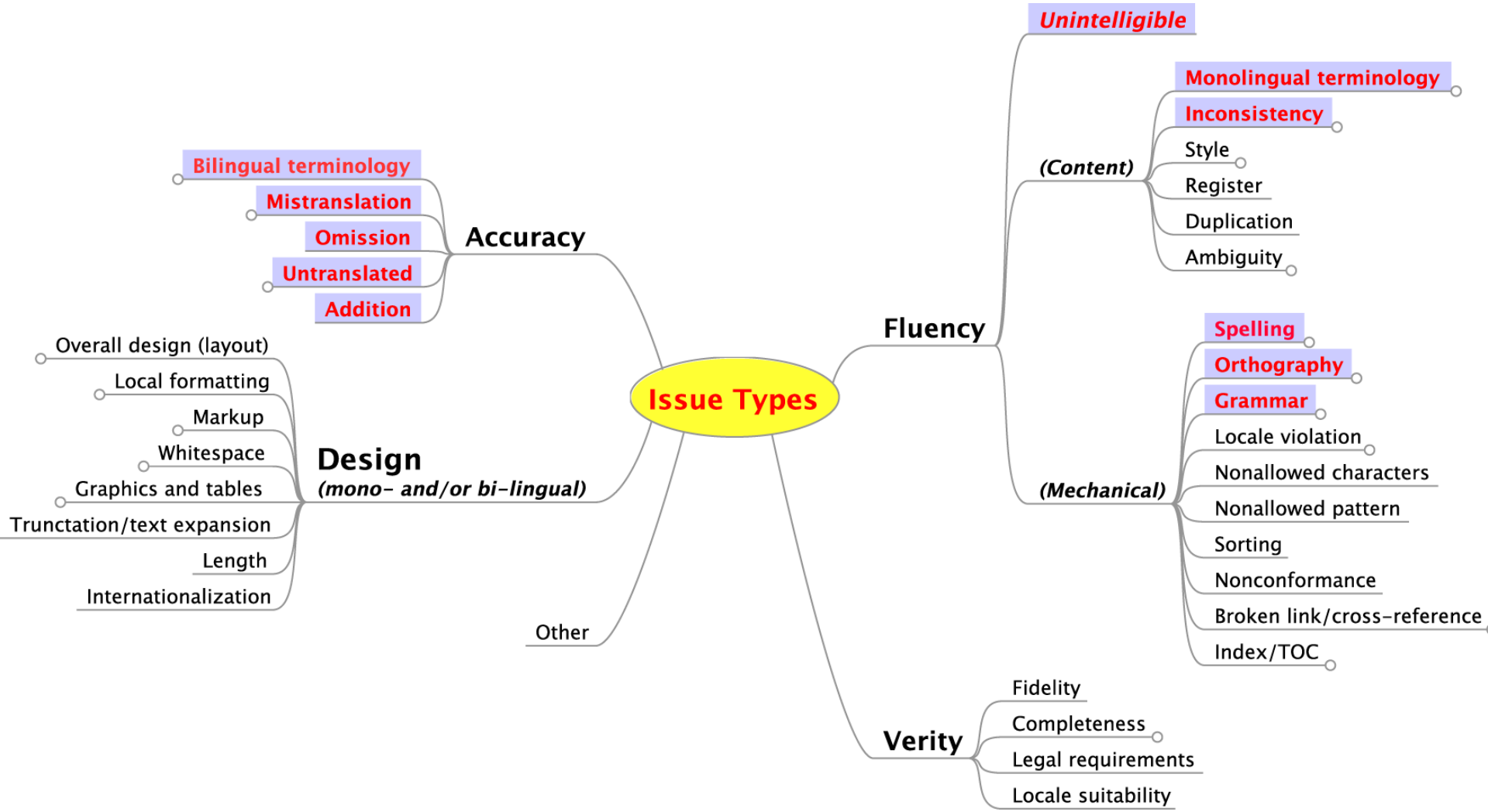
1. Lucia Specia
2. Eleftherios Avramidis
3. Carolina Scarton
4. José Guilherme Camargo de Souza
5. Matteo Negri
6. Jie Jiang
7. Mark Fishel
8. Niko Papula
9. Konstantinos Chatzitheodorou
10. Shadi Saleh
11. Christian Buck

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-c480dRoVugO6RtuvCoPrpayvbENYI/edit#heading=h.gjdgxs

MQM Error Categories



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$_{\text{miss}}$ ei sisalda$_{\text{lex}}$ uusi elemente$_{\text{lex}}$ tema$_{\text{lex}}$ hinnangu$_{\text{infl}}$ jaoks$_{\text{miss}}$
9::hyp-err-cats: komisjon leiab et ei looda$_{\text{lex}}$ uusi andmeid$_{\text{lex}}$ oma$_{\text{lex}}$ hinnang$_{\text{infl}}$

http://terra.cl.uzh.ch

Human Translation Error Corpora

1	language	context
Original context	fr	Le travailleur ressortissant d'un État membre occupé sur le territoire d'un autre État membre bénéficie de l'égalité de traitement en matière d'exercice des droits syndicaux, y compris le droit de vote et l'accès aux postes d'administration ou de direction d'une organisation syndicale; il peut être exclu de la participation à la gestion d'organismes de droit public et de l'exercice d'une fonction de droit public. Il bénéficie, en outre, du droit d'éligibilité aux organes de représentation des travailleurs dans l'entreprise. Les membres de la famille (conjoint, descendants de moins de 21 ans ou à charge et ascendants à charge) d'un travailleur employé sur le territoire d'un autre État membre ont le droit de s'y installer avec lui, quelle que soit leur nationalité.
Target sentence	en	The members of the family who are admitted into another Member State [under the title]$_{TR\text{-}SI\text{-}TL}$ of a " family group " are entitled to equal treatment , i.e.
Alternative student context	en	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 [right]$_{LA\text{-}IA\text{-}NU}$ in workers ' [representation]$_{LA\text{-}ST\text{-}AW}$ bodies of the company . The family members of any EU [worker 's]$_{LA\text{-}HY\text{-}PU}$ - [husband , wife]$_{LA\text{-}RE\text{-}IN}$, descendants aged under 21 years [old]$_{LA\text{-}ST\text{-}AW}$, [dependents]$_{LA\text{-}IA\text{-}NU}$ children and dependent [ascendants]$_{LA\text{-}TL\text{-}FC}$ - whatever their [nationality have]$_{LA\text{-}HY\text{-}PU}$ the right to settle with the worker in the host Member State .

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]

RMSE		
Dataset	**Baseline - mean of training**	**Baseline 17 features**
Turk Translator 1	0.1790	0.1669
Turk Translator 2	0.1986	0.1771
Turk Translator 3	0.2111	0.2084
Turk Translator 4	0.1756	0.1700
Turk Editor 1-1	0.1800	0.1692
Turk Editor 1-2	0.1844	0.1727
Turk Editor 1-3	0.1910	0.1867

Urdu-English turker translations

- Baseline features, predicting Meteor [0,1]

RMSE		
Dataset	**Baseline - mean of training**	**Baseline 17 features**
Turk Editor 2-1	0.1854	0.1760
Turk Editor 2-2	0.2005	0.1824
Turk Editor 2-3	0.2038	0.1876
Turk Editor 3-1	0.2027	0.1983
Turk Editor 3-2	0.2126	0.2077
Turk Editor 3-3	0.2103	0.2014
Turk Editor 4-1	0.1700	0.1660

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:

MAE		
Baseline - mean of training	**Baseline 17 features**	**17 + Discourse features**
0.3389	0.2324	**0.2219**

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)