HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation

Serge Gladkoff * & Lifeng Han **

lifeng.han@{manchester.ac.uk, adaptcentre.ie} serge.gladkoff@logrusglobal.com

* CEO at Logrus Global https://logrusglobal.com/
** ADAPT Research Centre, DCU, IE (former) & The University of Manchester, UK (Now)
@ LREC2022, June 20th, Marseille, France

Something bonus: our tutorial on MTE

https://github.com/poethan/LREC22_MetaEval_Tutorial

Meta-Evaluation of Translation Evaluation Methods: a systematic up-to-date overview

Lifeng Han * and Serge Gladkoff **

* The University of Manchester (*current*), UK & ADAPT Research Centre, DCU (*former*), Ireland
** Logrus Global (https://logrusglobal.com)

Half-day Tutorial @ LREC2022, June 20th, Marseille, France

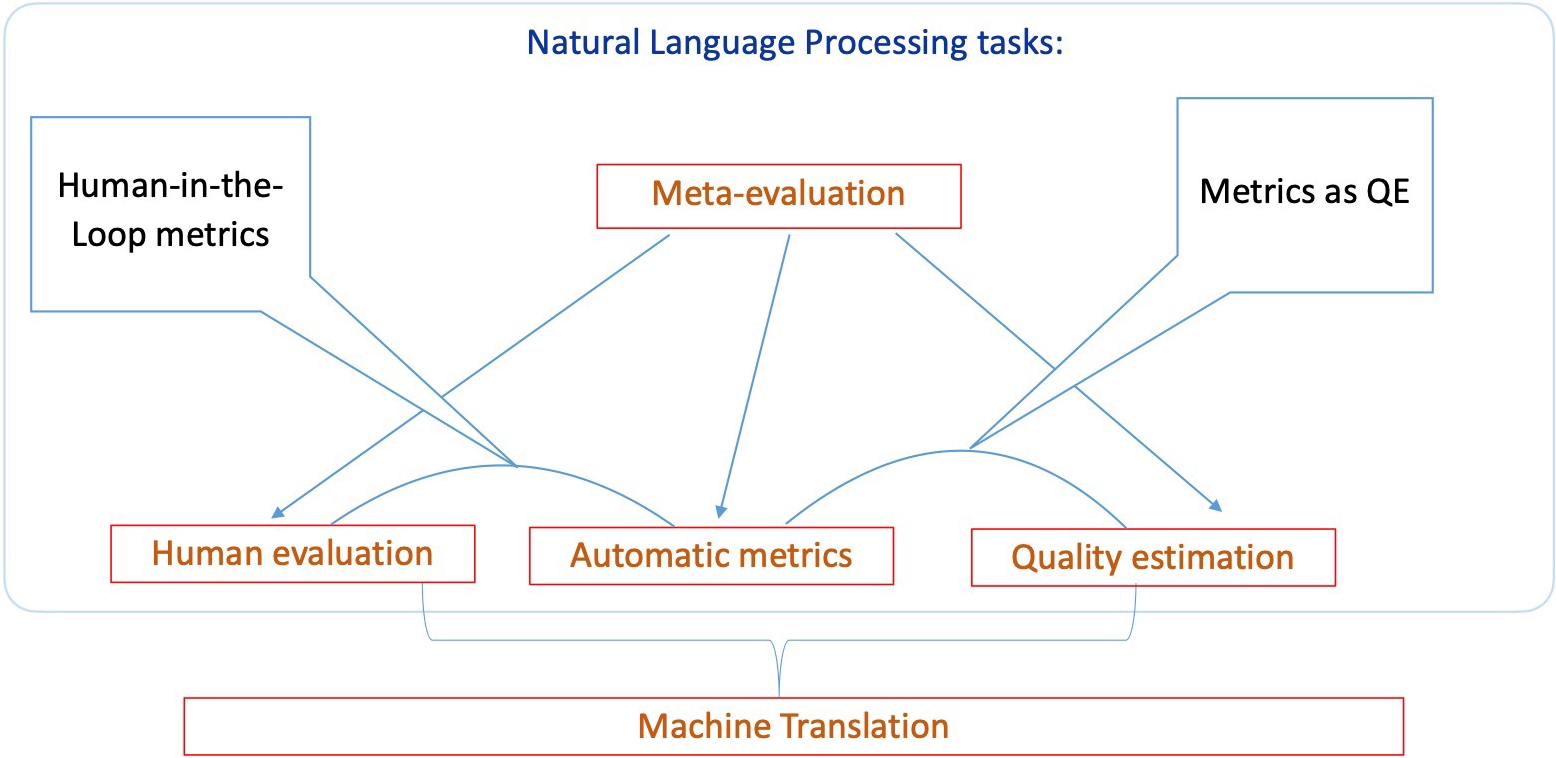
lifeng.han@{manchester.ac.uk, adaptcentre.ie} *serge.gladkoff@logrusglobal.com*

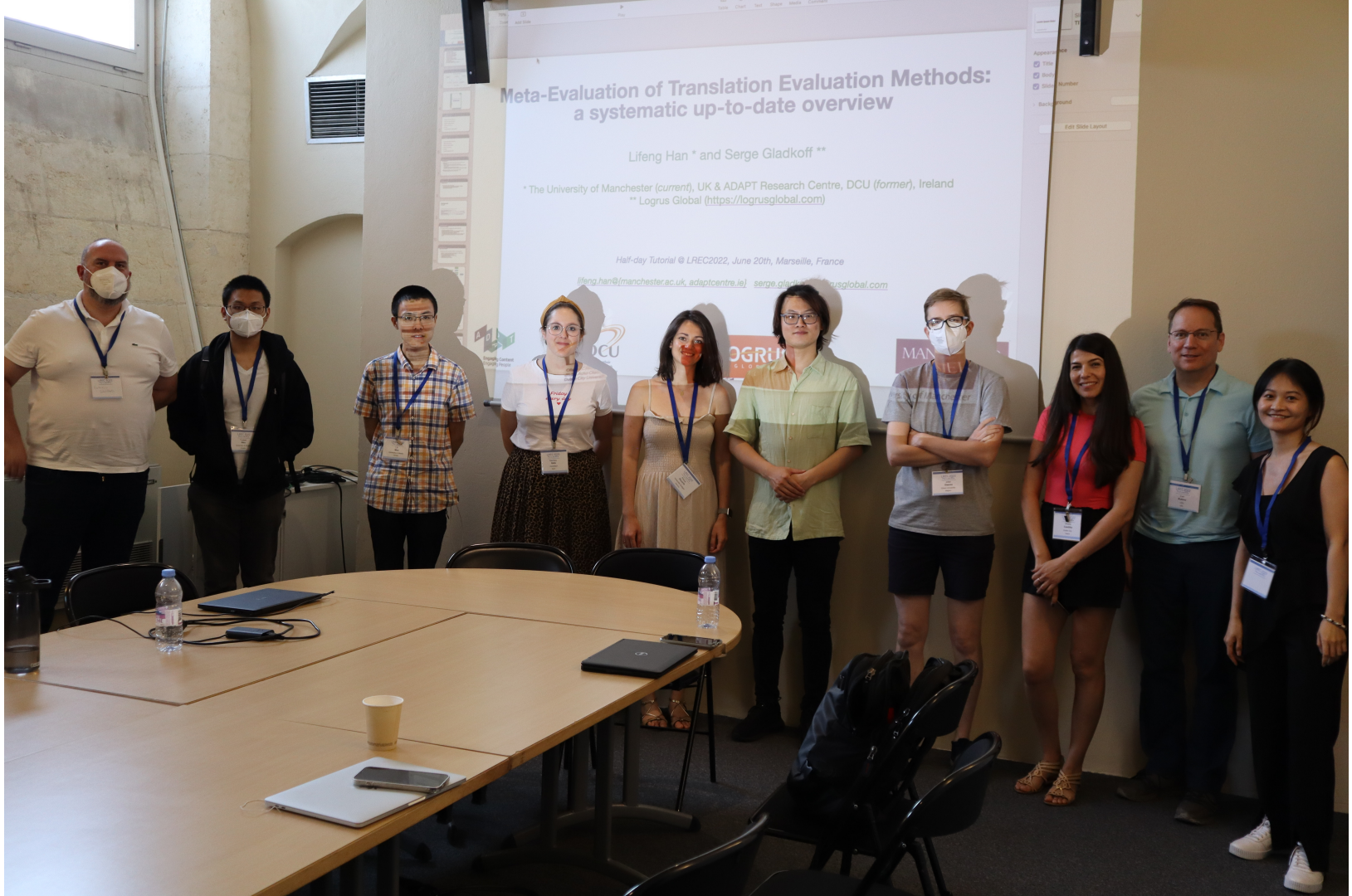
Engaging Content
Engaging People

Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

The University of Manchester

Content - structure/topics




Meta-Evaluation of Translation Evaluation Methods:
a systematic up-to-date overview
Lifeng Han * and Serge Gladkoff **
* The University of Manchester (current), UK & ADAPT Research Centre, DCU (former), Ireland
** Logrus Global (https://logrusglobal.com)
Half-day Tutorial @ LREC2022, June 20th, Marseille, France

Presentation schedule: **HOPE**

Motivation

Design: HOPE

Experimental evaluation

Summary

Motivation

Traditional automatic evaluation metrics for MT have been widely criticized by linguists due to their:

- low accuracy

- lack of transparency

- focus on language mechanics rather than semantics

- and low agreement with human quality evaluation.

Human evaluations in the form of MQM-like scorecards have always been carried out in real industry setting by both clients and translation service providers (TSPs). However:

- traditional human translation quality evaluations are costly to perform

- and go into great linguistic detail, raise issues as to inter-rater reliability (IRR)

- and are not designed to measure quality of worse than premium quality translations.

Motivation

For large-scale deployment of MT, a more appropriate quality metric is required which:

a) allows for faster learning curve for evaluators to be applied correctly;

b) is faster to apply;

c) is specifically designed to address less than perfect MT output of “good enough quality”;

d) does not track so many unnecessary linguistic details as standard MQM metrics, designed as a tool to measure near-premium quality of human translations.

Design

HOPE: a task-oriented and **H**uman-centric evaluation framework for machine translation output based **o**n professional **P**ost-**e**diting annotations.

It contains:

- a limited number of commonly occurring error types (proper name, impact, required adaptation, terminology, grammar, accuracy, style, and proofreading error)
- uses a scoring model with geometric progression of error penalty points (EPPs) reflecting error severity level to each translation unit.
- Error annotation and scoring can be done either without post-editing itself, or during post-editing towards a newly generated post-edited reference translation

Design: factors

Code	Definition	Explanation
IMP	Impact	The translation fails to convey the main thought clearly (even if translation may be literally correct, but proper translation should not be literal in target language, or has poor expression of the main thought).
RAM	Required Adaptation is Missing	Source contains error that has to be corrected, or target market requires substantial adaptation of the source, which translator failed to make. Impact on end user suffers.
TRM	Terminology	Incorrect terminology, inconsistency of translation of entities (forms, sections, etc.)
UGR	Ungrammatical	Translation is ungrammatical - needs to be fixed to convey the meaning properly.
MIS	Mistranslation	Translation distorts the meaning of the source, and presents mistranslation or accuracy error.
STL	Style	Translation has poor style, but is not necessarily ungrammatical or formally incorrect.
PRF	Proofreading error	Linguistic error which does not affect accuracy or meaning transfer, but needs to be fixed.
PRN	Proper Name	A proper name is translated incorrectly.

Design: scoring segments

Errors of each type can have the following severity differences: (minor, medium, major, severe, critical) with the corresponding values (1, 2, 4, 8, 16).

Error points for each Translation Unit (TU) are added to form the Error Point Penalty (EPP) of the TU (EPPTU) under-study.

Each TU has its own EPPTU not depending on other TUs.

Importantly, repeated errors in different TUs are not counted as one error.

The system-level score of HOPE is calculated by the sum of overall segment-level EPPTUs.

Design: scoring HOPE

$$\text{EPPTU} = \sum_{i} Error_i \times Severity(i)$$

$$\text{HOPE} = \sum_{TU_j} EPPTU_j = \sum_{i,j} Error_i \times Severity(i)$$

Design: segment/sentence & word-level

segment/sentence-level: We apply this metric into a sentence level (or segment-level) error severity classification, i.e. **minor** *vs* **major** with the EPPTU score (1~4) *vs* 5+.

The benefit of such design is that it immediately allows to distill sentences with only minor errors, with EPPTUs 1, 2, 3 and 4, and sentences with major errors (EPPTUs 5 and more).

One can say that EPPTU 1-4 is precisely what is often meant by ``good enough quality'' of MT, where budget, time or frequency of visiting the content does not provide for premium quality translation and lower quality is just fine.

Design: segment/sentence & word-level

The word level HOPE: follows the segment-level indicators including “unchanged”, “good enough”, and “must be fixed”.

However, the statistics will be reflected at word level, e.g. how many words of the whole document/text belong to each of the three categories.

Both segment/sentence-level and word-level HOPE indicators can be used to reflect the overall MT quality in translating the overall material/document. However, they can tell different aspects of the MT systems, e.g. when there are many sentences falling into very different length (short {vs} long sentences).

Experimental evaluation

The pilot experiments contain two tasks.

Task-I is carried out using:

- English=>Russian (EN=>RU) language pair from marketing document in technical domain with 111 sentences (segments), using two MT engines, Google Translator and a customised MT engine.

Task-II uses a document from business domain containing 671 segments (3,339 words) on the same translation direction but using an alternative NMT engine DeepL.

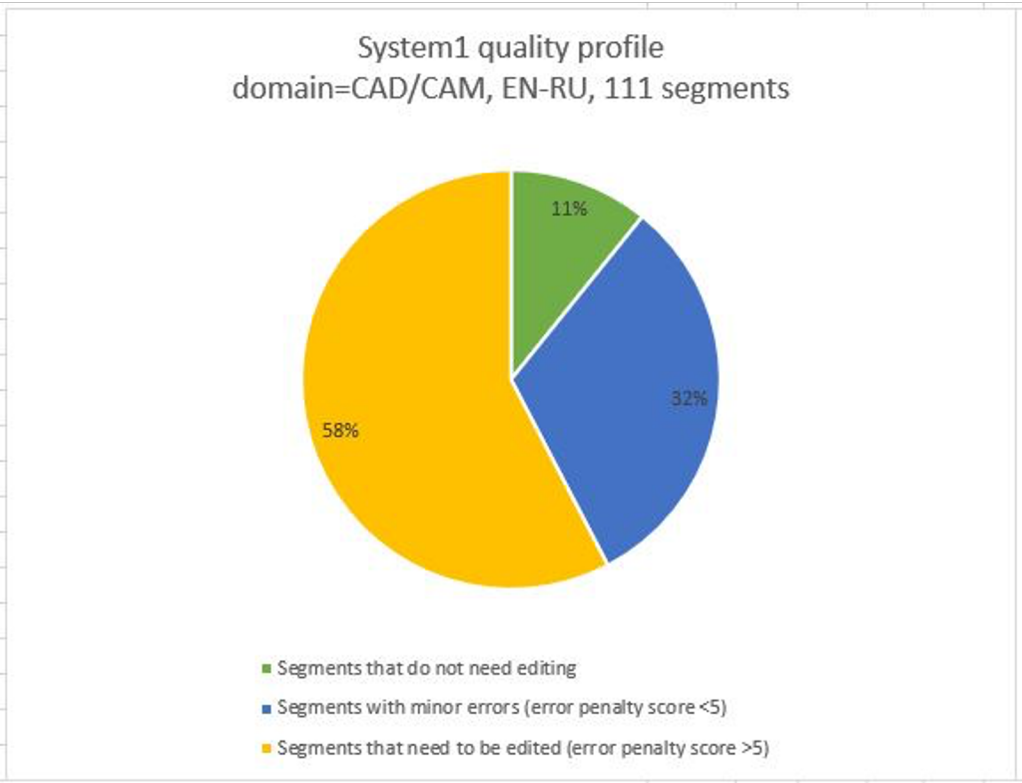
Experimental evaluation: sentence level

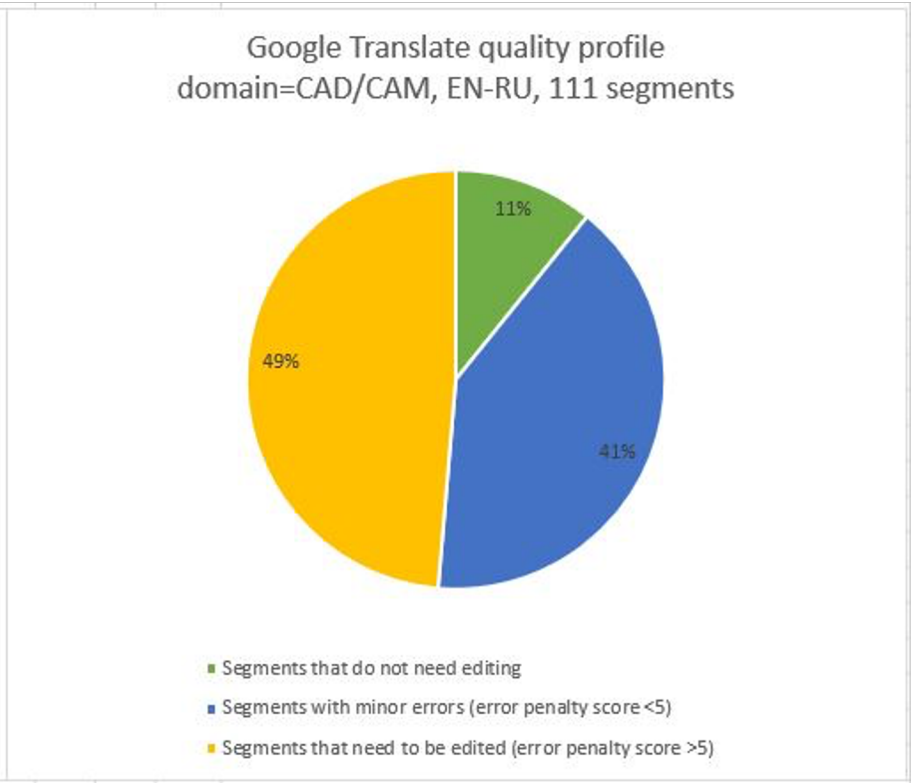
A	B	C	D	E	F	G	H	I	J	K
SRC	System1	NOC	PRN	ACR	STL	TRM	IMP	UGR	PRF	SEGS
TOTAL by 111 segments		12	32	168	192	235	80	20	8	735
% of total 111 segments	Segments that do not need editing	**11%**	**4%**	**23%**	**26%**	**32%**	**11%**	**3%**	**1%**	**6.6**
TOTAL of segments with scores 1,2,3 and 4		**35**								
% of segments with scores 1,2,3 and 4	Segments with minor errors (error penalty score <5)	**32%**								
TOTAL of segments with scores >4		**64**								
	Segments that need to be edited (error penalty score >5)	**58%**								

A	B	M	N	O	P	Q	R	S	T	U
SRC	Google MT	NOC	PRN	ACR	STL	TRM	IMP	UGR	PRF	SEGS
TOTAL by 111 segments		12	22	164	205	207	58	16	6	678
% of total 111 segments	Segments that do not need editing	**11%**	**3%**	**24%**	**30%**	**31%**	**9%**	**2%**	**1%**	**6.1**
TOTAL of segments with scores 1,2,3 and 4		**45**								
% of segments with scores 1,2,3 and 4	Segments with minor errors (error penalty score <5)	**41%**								
TOTAL of segments with scores >4		**54**								
	Segments that need to be edited (error penalty score >5)	**49%**								

Each error categories in % and overall how many percents of sentences fall into no-change/minor/major-edit

Experimental evaluation: comparison of two quality profiles





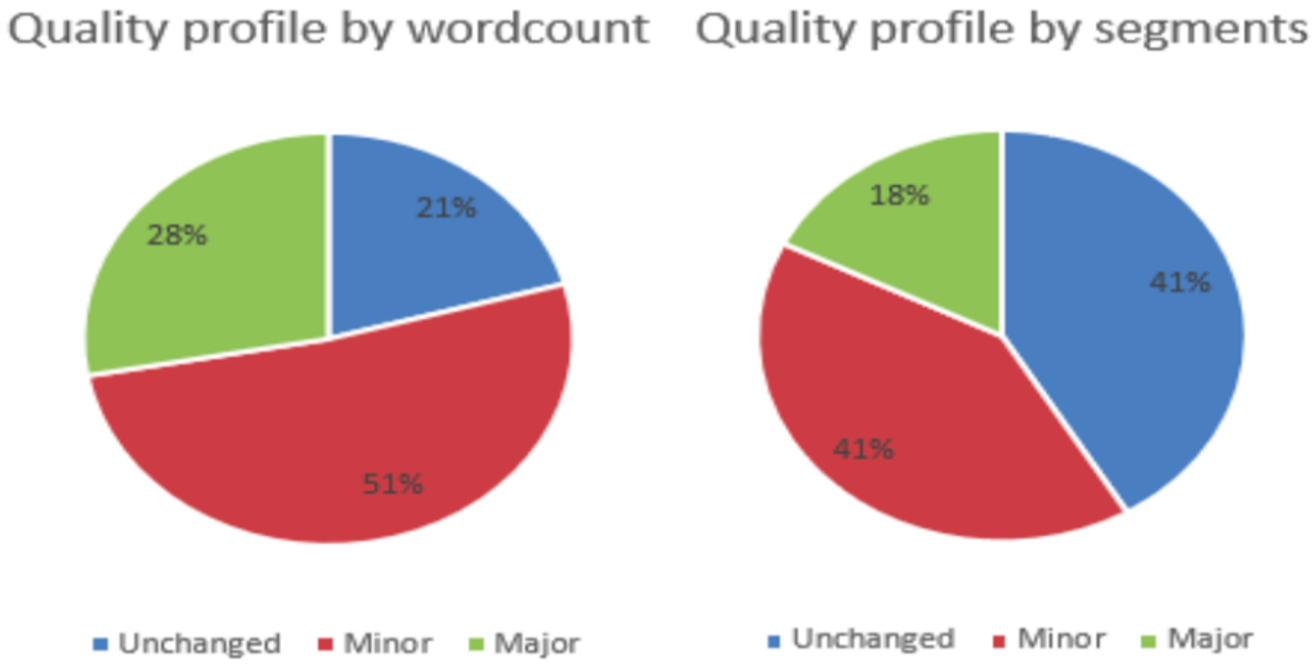
Task-I HOPE quality indicators: comparison of System1 vs Google Translate

Experimental evaluation: word level *vs* segment level

	Unchanged	Minor	Major
Total Segments (671)	278	275	118
Percent of Segments	41%	41%	18%
Wordcount (3339)	691	1718	930
Percent of Wordcount	21%	51%	28%

Quality Indicators via Segment {vs} Word Level HOPE: number counts and percentage

Experimental evaluation: word vs segment



HOPE quality indicators via word-level (left) {vs} segment-level (right) in percentage

Summary

We designed HOPE metric:

can be seen as an MQM implementation tailored specifically for evaluating MT output.

The approach has several key advantages:

- ability to measure and compare less than perfect MT output from different systems
- ability to indicate human perception of quality
- immediate estimation of the labor effort required to bring MT output to premium quality / good enough
- low-cost and faster application, as well as higher IRR

Initial experimental eval on: segment/word-level HOPE using en=>ru works!

Source files (corpus and scoring): https://github.com/lHan87/HOPE

Demo: HOPE

- https://github.com/poethan/LREC22_MetaEval_Tutorial (downloading link)

- A Link to **download our tutorial** draft PPT slides (will be updated after tutorial)
- Or this drive:
- https://drive.google.com/drive/folders/1sajkcrnDgTOFiYVkFjYqh9EDYUhneYqA?usp=sharing

References (selected)

Gladkoff, S., Sorokina, I., Han, L., and Alekseeva, A. (2021). Measuring uncertainty in translation quality evaluation (TQE). CoRR, abs/2111.07699.

Han, L., Jones, G., and Smeaton, A. (2020b). Mul-tiMWE: Building a multi-lingual multi-word expression (MWE) parallel corpora. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 2970–2979, Marseille, France, May. European Language Resources Association.

Han, L., Smeaton, A., and Jones, G. (2021a). Translation quality assessment: A brief survey on manual and automatic methods. In Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age, pages 15–33, online, May. Association for Computational Linguistics.

Han, L., Sorokina, I., Erofeev, G., and Gladkoff, S. (2021b). cushLEPOR: customising hLEPOR metric using optuna for higher agreement with human judgments or pre-trained language model labse. In Proceedings of Six Conference on Machine Translation (WMT2021), In Press. Association for Computational Linguistics.

Han, L. (2022). An Investigation into Multi-Word Expressions in Machine Translation. PhD thesis, Dublin City University. https://doras.dcu.ie/26559/.

Lommel, A., Burchardt, A., and Uszkoreit, H. (2014). Multidimensional quality metrics (mqm): A frame-work for declaring and describing translation quality metrics. Traduma`tica: tecnologies de la traduccio´, 0:455–463, 12.