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

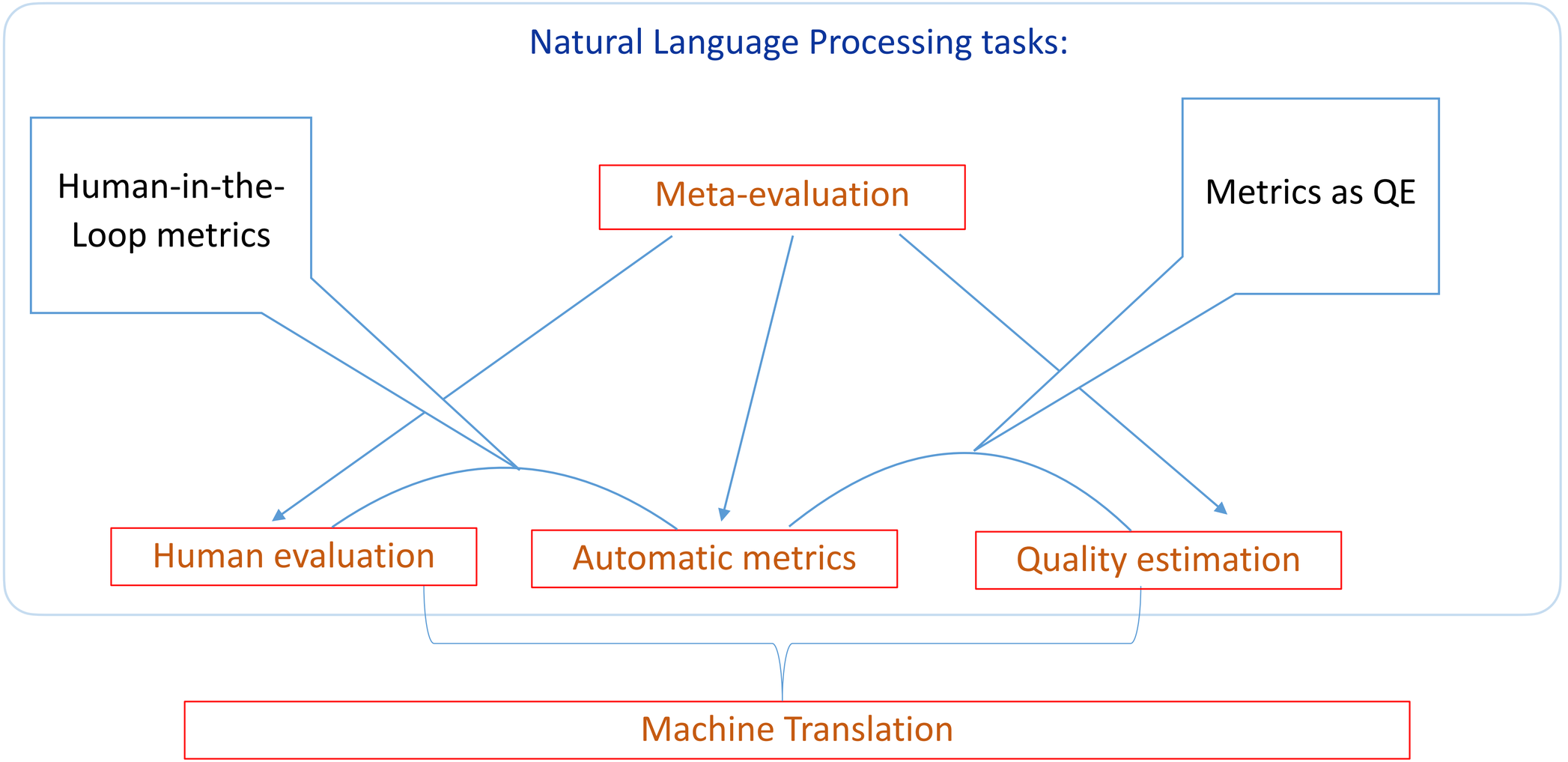
Ollscoil Chathair
Bhaile Átha Cliath
Dublin City University

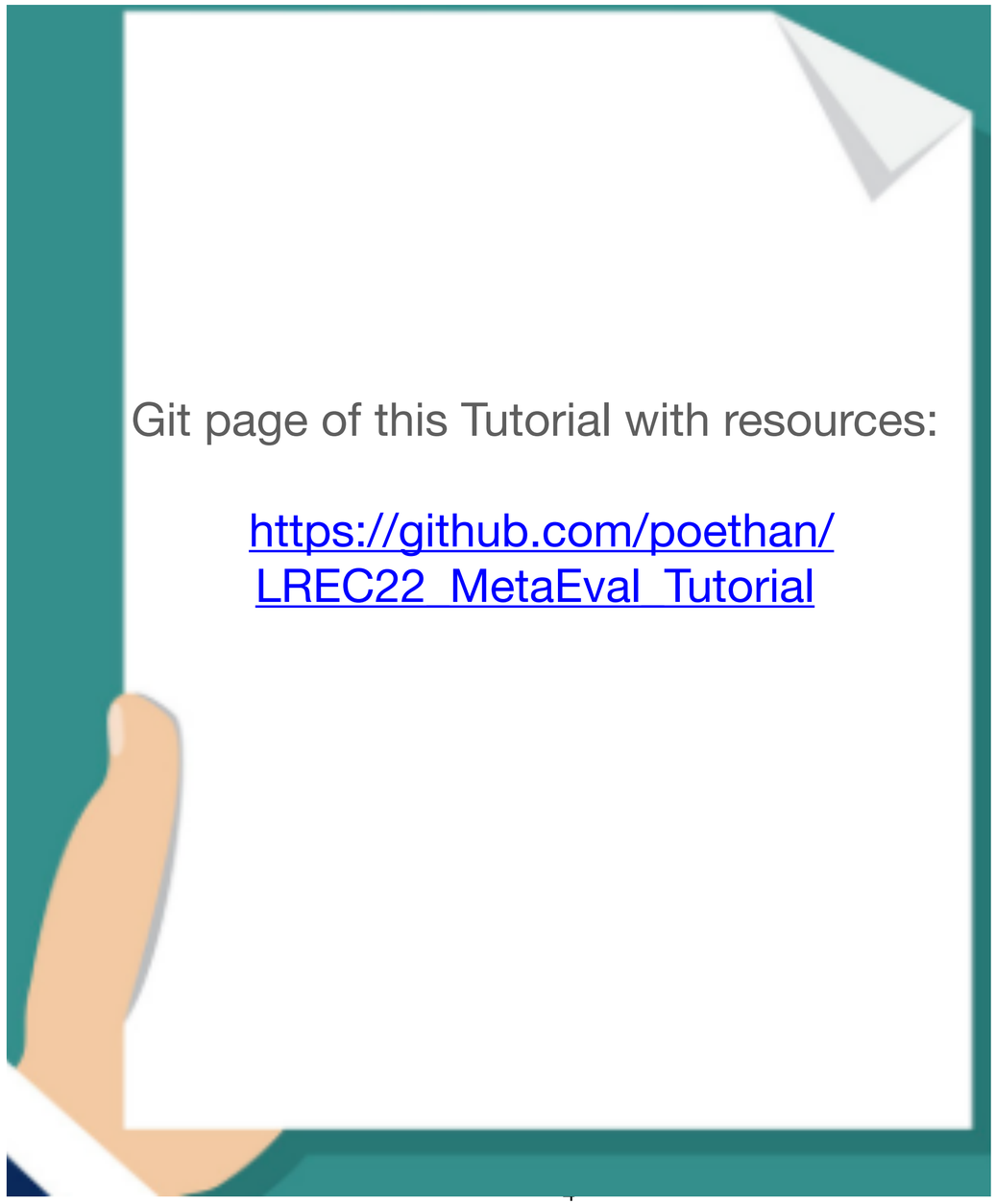
The University of Manchester

Abbreviations

- MT: machine translation
- SMT: statistical machine translation
- NMT: neural machine translation
- MTE: machine translation evaluation
- TQA: translation quality assessment
- TQE: translation quality evaluation
- QE: quality estimation
- HumanEval (HE): human evaluation
- AutoEval (AE): automatic evaluation
- MetaEval: meta-evaluation (evaluating the evaluation methods)
- NE: named entities
- MWE: multi-word expression
- EN/PL/DE/ZH: English/Polish/German/Chinese

Content - structure/topics




Git page of this Tutorial with resources:
https://github.com/poethan/
LREC22_MetaEval_Tutorial

Content - list

- Background (& motivation of this work)
- Related work (surveys and overviews)
- HumanEval, AutoEval, MetaEval
- Discussion and Perspectives
- Conclusions
- => Appendices (references, codes, platforms)

Content

- **Background (& motivation of this work)**
- Related work (earlier surveys)
- HumanEval, AutoEval, MetaEval
- Discussion and Perspectives
- Conclusions
- => Appendices (evaluating TQA, QE)

Background

Driven factors

- (M)TE as a key point in (M)Translation Quality Assessment & MT model development
  - Human evaluations as the golden criteria of assessment
  - Automatic metrics for MT system tuning, parameter optimisation, Easy to use for scoring.
  - => however, criticisms exist from both above two
  - New MTE methods needed for distinguishing high performance MT systems

- (M)TE applications/influences in other NLP evaluation tasks
  - Sumarisation, generation, image captioning, etc.

Background

Driven factors

- Related earlier surveys and overviews:
  - not covering recent years development (EuroMatrix07, GALE09)
  - Specialised/focused to certain domain (Secară 05)
  - or very initial, brief (Goutte 06)

Background

Our work

This tutorial was initially derived from our earlier pre-print: L. Han (2016)

Machine Translation Evaluation Resources and Methods: A Survey
https://arxiv.org/abs/1605.04515 updated-2018
Lifeng Han (2018) Machine Translation Evaluation Resources and Methods: A Survey https://arxiv.org/abs/1605.04515v8
Han et al. (2021) Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. https://aclanthology.org/2021.motra-1.3/
Lifeng Han (2022) An Overview on Machine Translation Evaluation. https://arxiv.org/abs/2202.11027 (in Chinese, English update forthcoming)

Background

aim: what we wanted

- a concise but structured classification and overview of MTE/TQA

  - Introduction to concepts in this field
  - extending related literature, e.g. recent years published MTE work/trend
  - Covering broader topics: HumanEval, AutoEval, MetaEval
  - easy to grasp what reader/researcher might need
  - - find their corresponding evaluation methods efficiently from existing ones
  - - or make their own that can be inspired by the methods this overview covers

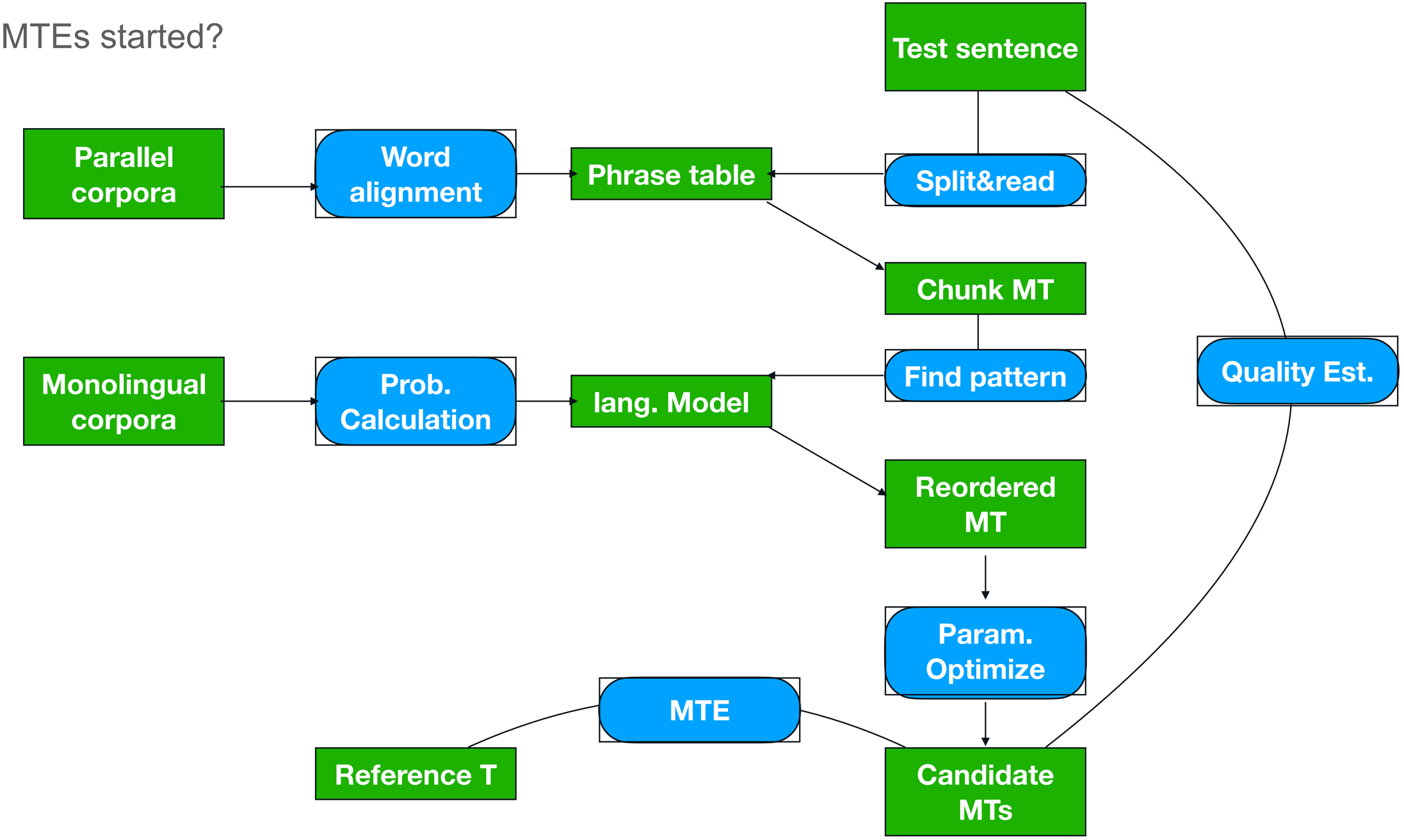
Background

aim: what we wanted

- We expect this tutorial to be helpful for different NLP task (Evaluations)

  - different NLP tasks share similarities and Translation is a relatively larger task that covers/impacts/interacts with others, e.g.
  - - text summarisation (Bhandari et al 2021),
  - - language generation (Novikova et al. 2017emnlp),
  - - searching (Liu et al. 2021acm),
  - - code generation (Liguori et al 2021)

Where MTEs started?



Ref: Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

Background - examples

TQA1: src + ref + output
(HE)

TQA2: src + output
(HE, QE)

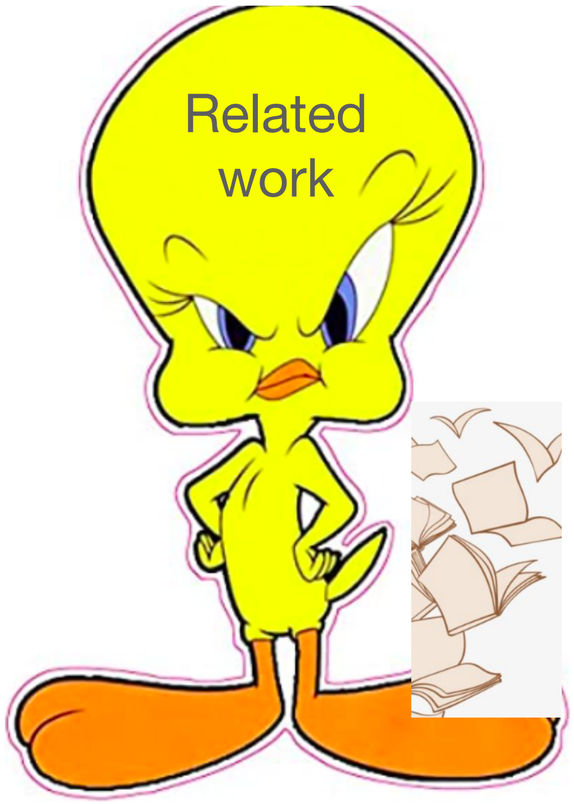
TQA3: ref + output
(HE, Metrics)

ZH source:	年年歲歲花相似，歲歲年年人不同
ZH pinyin:	Nián nián suì suì huā xiāng sì, suì suì nián nián rén bù tóng.
EN reference:	The flowers are similar each year, while people are changing every year.
EN MT output:	One year spent similar, each year is different

Ref: Han, Lifeng, Gareth, Jones and Alan F., Smeaton (2020) MultiMWE: building a multi-lingual multi-word expression (MWE) parallel corpora. In: 12th International Conference on Language Resources and Evaluation (LREC), 11-16 May, 2020, Marseille, France. (Virtual).

Content

- Background (& motivation of this work)
- **Related work (earlier surveys)**
- HumanEval, AutoEval, MetaEval
- Discussion and Perspectives
- Conclusions
- => Appendices (references, codes, platforms)



cartoon, https://www.amazon.ca/Tweety-Bird-not-happy-Decal/

Related work

- earlier surveys on TQA/MTE

- Alina Secară (2005): in Proceedings of the eCoLoRe/MeLLANGE workshop. page 39-44.
  - A special focus on error classification schemes
  - from translation industry and translation teaching institutions
  - a consistent and systematic error classification survey till early 2000s

- Cyril Goutte (2006): Automatic Evaluation of Machine Translation Quality. 5 pages.
  - surveyed two kind of metrics to date (2006): string match, and IR inspired.
  - String match: WER, PER;
  - IR style technique (n-gram precision): BLEU, NIST, F, METEOR

Related work

- earlier surveys on TQA/MTE

- EuroMatrix (2007): 1.3: Survey of Machine Translation Evaluation. In EuroMatrix Project Report, Statistical and Hybrid MT between All European Languages, co-ordinator: Prof. Hans Uszkoreit

  - Human evaluation and automatic metrics for MT, to date (2007).
  - An extensive referenced metric lists, compared to other earlier peer work.

Related work

- earlier surveys on TQA/MTE

- Bonnie Dorr (2009) edited DARPA GALE program report. Chap 5: Machine Translation Evaluation.

  - Global Autonomous Language Exploitation programme from NIST.
  - Cover HE (*intro*) and Metrics (*heavier*)
  - Finishing up with METEOR, TER-P, SEPIA (a syntax aware metric, 2007), etc.
  - Concluding the limitations of their metrics: high resource requirements, some seconds to score one segment pair, etc. Calls for better automatic metrics

Related work

- earlier surveys on TQA/MTE

- Màrquez L. Dialogue 2013 invited talk, extended. automatic evaluation of machine translation quality. @EAMT
  - An emphasis on linguistically motivated measures
  - introducing Asiya online interface by UPC for MT output error analysis
  - Introducing QE (fresh TQA task in WMT community)

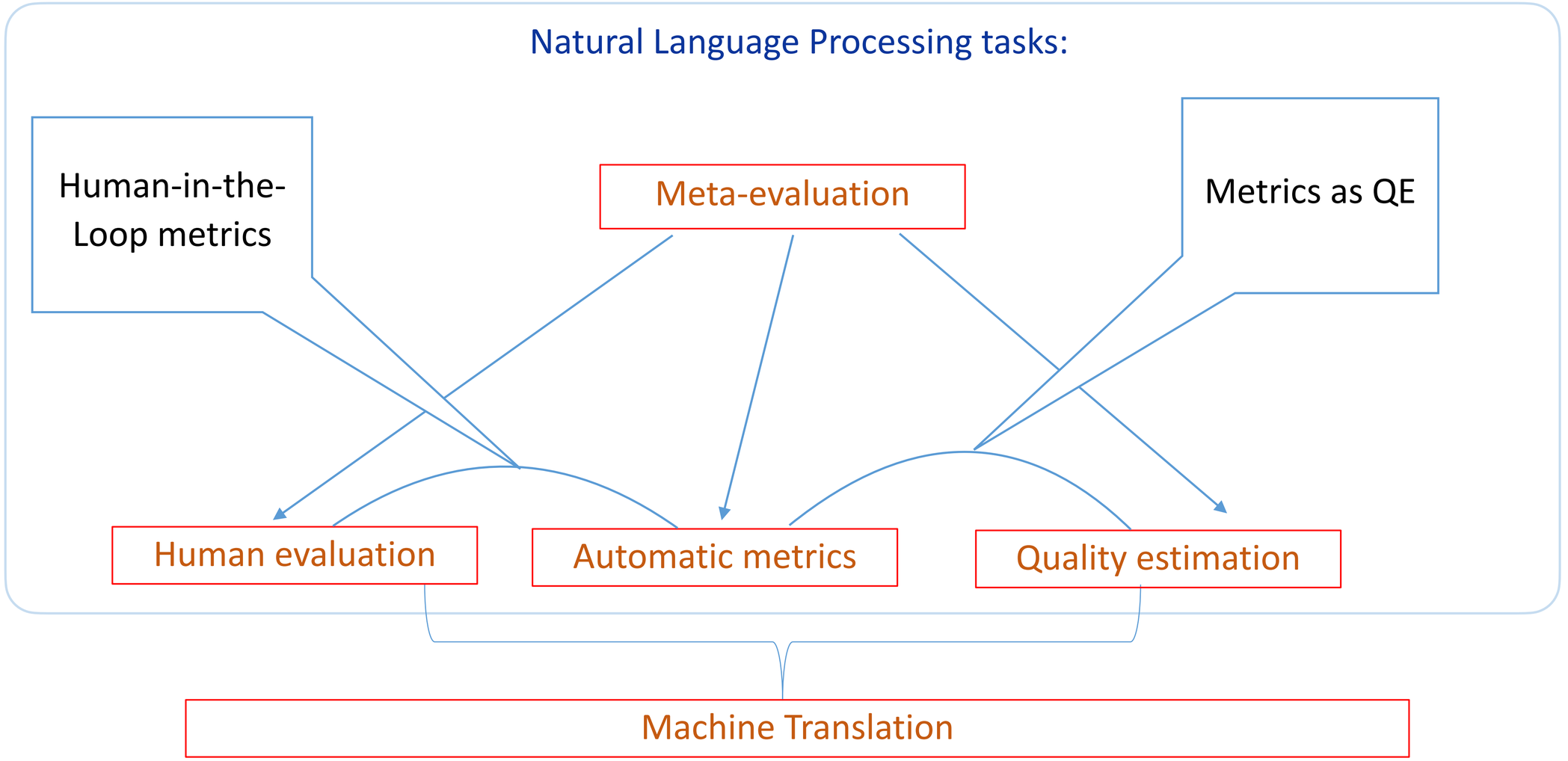
- Our *pre-prints and conference paper*:
- Han and Wong (2016): Machine Translation Evaluation: A Survey https://arxiv.org/abs/1605.04515 (updated in 2018 ->)
  - Han (2018) Machine Translation Evaluation Resources and Methods: A Survey https://arxiv.org/abs/1605.04515v8
  - Han, Smeaton, and Jones (2021) Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. https://aclanthology.org/2021.motra-1.3/
  - Lifeng Han (2022) An Overview on Machine Translation Evaluation. https://arxiv.org/abs/2202.11027 (in Chinese)

Content - systematic overview

- Background (& motivation of this work)
- Related work (earlier surveys)
- **HumanEval, AutoEval, MetaEval**
- Discussion and Perspectives
- Conclusions
- => Appendices (references, codes, platforms)

cartoon, https://www.bbc.com/news/

Looking back - the structure



Our classifications - MTE paradigm

Manual methods

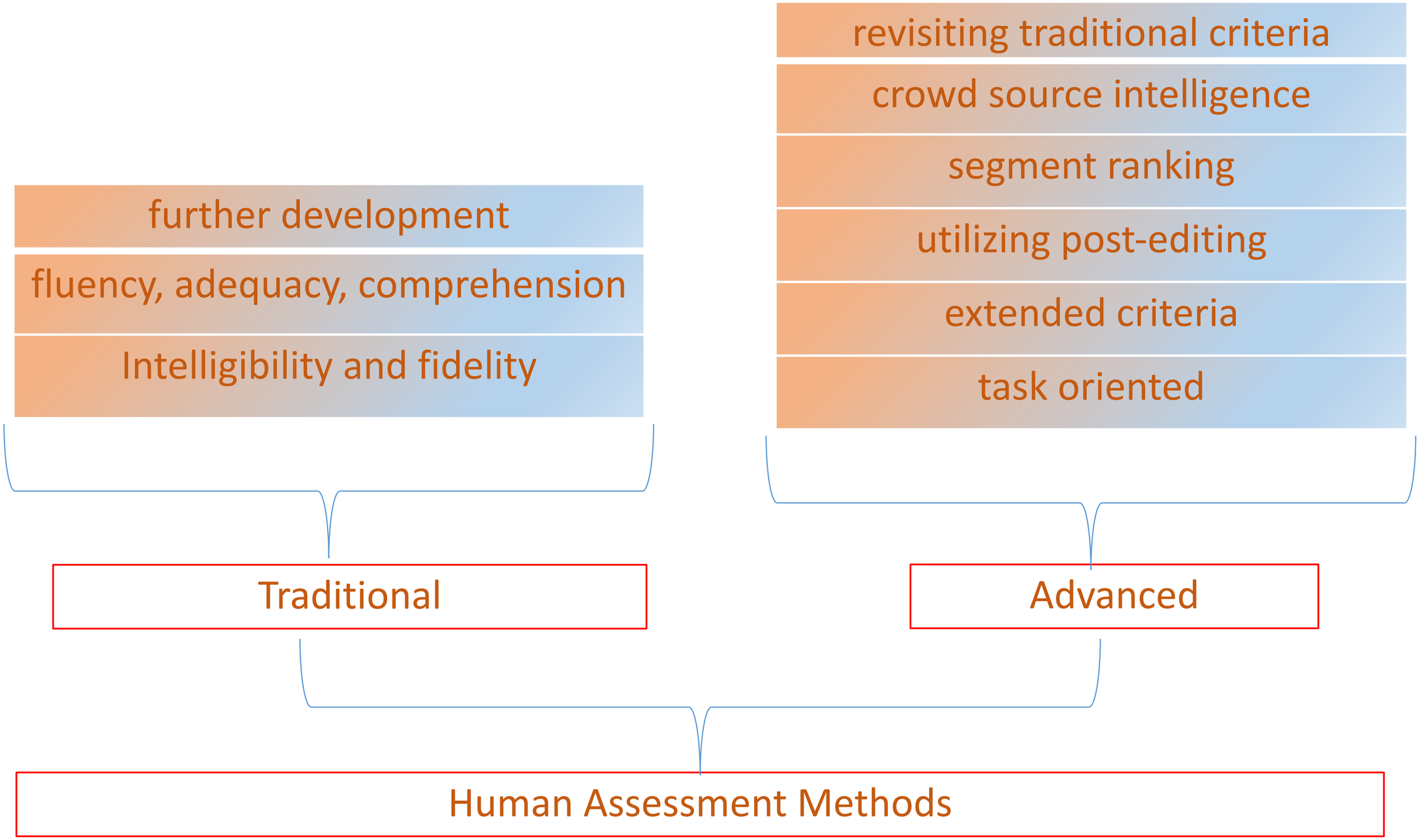
- started since Translation was deployed
- a boost development with together MT technology
- traditional criteria (from 1960s) *vs* advanced (further developed)

Automatic methods - 2000s on when MT making sense & popular

- Metrics (traditional): word n-gram match, linguistic features, statistical models, then ML/DL based.
- QEs (later developed): without relying on reference translations.

Evaluating MTE methods (meta-evaluation)

- statistical significance, agreement, correlations, test set, size, etc.



Ref: Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

Intelligibility and Fidelity

The earliest human assessment methods for MT can be traced back to around 1966. They include the intelligibility and fidelity used by the automatic language processing advisory committee (ALPAC) [38]. The requirement that a translation is intelligible means that, as far as possible, the translation should

read like normal, well-edited prose, and be readily understandable in the same way that such a translation would be understandable if originally composed in the translation language. The requirement that a translation is of high fidelity or accuracy includes the requirement that the translation should, as little as possible, twist, distort, or controvert the meaning intended by the original.

Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

In the 1990s, the Advanced Research Projects Agency (ARPA) created a methodology to evaluate MT systems using adequacy, fluency and comprehension of MT output [50], which was subsequently adapted for use in MT evaluation campaigns including [284].

To set up this methodology, a human assessor is asked to look at each fragment, delimited by syntactic constituents and containing sufficient information, and judge its adequacy on a scale of 1-to-5. Results are computed by averaging the judgements over all of the decisions in the translation set.

Fluency evaluation is compiled in the same manner as for adequacy except that the assessor is asked to make intuitive judgements on a sentence-by-sentence basis for each translation. Human assessors are asked to determine whether the translation is good English without reference to the correct translation. Fluency evaluation determines whether a sentence is well-formed and fluent in context.

Comprehension relates to "Informativeness", whose objective is to measure a system's ability to produce a translation that conveys sufficient information, such that people can gain necessary information from it. The reference set of expert translations is used to create six questions with six possible answers respectively including, "none of the above" and "cannot be determined".

Adequacy: the quantity of the information existent in the original text that a translation contains

Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

Further Development

In work by Bangalore *et al.* [13] the authors classified accuracy into several categories, including simple string accuracy, generation string accuracy, and two corresponding tree-based accuracy. Other work in [229] found a correlation between fluency and the number of words it takes to distinguish between human translation and MT output.

The Linguistics Data Consortium (LDC) designed two five-point scales representing fluency and adequacy for the annual NIST MT evaluation workshop (referring to *LDC2003T17*) [9]. The developed scales became a widely used methodology when manually evaluating MT by assigning values. The five point scale for adequacy indicates how much of the meaning expressed in the reference translation is also expressed in a translation hypothesis (from None to All); the second five point scale indicates how fluent the translation is, involving both grammatical correctness and idiomatic word choices (from Incomprehensible to Flawless English).

Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

HE/Human TQA - categories

- task oriented: in light of the tasks for which their output might be used, e.g. good enough MT (Gladkoff and Han, 2022)

- Extended Criteria: suitability, interoperability, reliability, usability, efficiency, maintainability, portability. (focus on MT system)

- Utilising Post-editing: compare the post-edited correct translation to the original MT output, the number of editing steps (TER, HTER, PER, HOPE)

- Segment Ranking: assessors were frequently asked to provide a complete ranking over all the candidate translations of the same source segment, WMTs.

- crowd-source intelligence: with Amazon MTurk, using fluency criteria, different scales set-up, with some quality control solutions. (Graham et al. 2013)

- Revisiting Traditional Criteria: ask the assessors to mark whatever errors from the candidate translations, with questions on adequacy and comprehension (En->Croatian/Serbian, Popovic 2021)

- sentence vs document level HumanEval

- eval on MT bottleneck MWEs (Han, Jones, Smeaton, 2020)

Extended Criteria

[139] extended a large range of manual evaluation methods for MT systems which, in addition to the earlier mentioned accuracy, include: *suitability*, whether even accurate results are suitable in the particular context in which the system is to be used; *interoperability*, whether it will operate seamlessly with other software or hardware platforms; *reliability*, i.e., does not break down or take a long time to get running again after breaking down; *usability*, easy to understand the interfaces, easy to learn and operate, and looks well; *efficiency*, when needed, keep up with the flow of dealt documents; *maintainability*, being able to modify the system in order to adapt it to particular users; and *porta-*

bility, one version of a system can be replaced by a new version, because MT systems are rarely static and tend to improve over time as resources grow and bugs are fixed.

Editing distance / Post-editing based

3.1.1 *Edit Distance*

By calculating the minimum number of editing steps to transform output to reference, (Su et al., 1992) introduce the word error rate (WER) metric into MT evaluation. This metric takes word order into account, and the operations include insertion (adding word), deletion (dropping word) and replacement (or substitution, replace one word with another), the minimum number of editing steps needed to match two sequences.

$$\text{WER} = \frac{\text{substitution+insertion+deletion}}{\text{reference}_{\text{length}}}. \quad (5)$$

However, due to the diversity of language expression, some so-called "wrong" order sentences by WER also prove to be good translations. To address this problem, the position-independent word error rate (PER) (Tillmann et al., 1997) is designed to ignore word order when matching output and reference. Without taking into account of the word order, PER counts the number of times that identical words appear in both sentences. Depending on whether the translated sentence is longer or shorter than the reference translation, the rest of the words are either insertion or deletion ones.

$$\text{PER} = 1 - \frac{\text{correc} - \max(0, \text{output}_{\text{length}} - \text{reference}_{\text{length}})}{\text{reference}_{\text{length}}}.$$

Lifeng Han (2018) Machine Translation Evaluation Resources and Methods: A Survey https://arxiv.org/abs/1605.04515v8

Another way to overcome the unconscionable penalty on word order in the Levenshtein distance is adding a novel editing step that allows the movement of word sequences from one part of the output to another. This is something a human post-editor would do with the cut-and-paste function of a word processor. In this light, (Snover et al., 2006) design the translation edit rate (TER) metric that adds block movement (jumping action) as an editing step.The shift option performs on a contiguous sequence of words within the output sentence. The TER score is calculated as:

$$\text{TER} = \frac{\#\text{of edit}}{\#\text{of average reference words}} \quad (7)$$

For the edits, the cost of the block movement, any number of continuous words and any distance, is equal to that of the single word operation, such as insertion, deletion and substitution.

However, it is not clear how to link the number of insertion, deletion and substitution to exactly how good is the quality of MT output.

How about there is no reference available? create a new reference produced by post-editing?

Lifeng Han (2018) Machine Translation Evaluation Resources and Methods: A Survey https://arxiv.org/abs/1605.04515v8

Utilising Post-editing

One alternative method to assess MT quality is to compare the post-editing correction to the original MT output. This type of evaluation is, however, time consuming and depends on the skills of the human assessor and post-editing performer. One example of a metric that is designed in such a manner is the human-targeted translation edit rate (HTER) [251] which is based on the translation edit rate (TER) metric [212] using the number of editing steps. Here, a human assessor has to find the minimum number of insertions, deletions, substitutions, and shifts to convert the system output into an acceptable translation. HTER calculates the minimum of edits to a *new targeted reference*, i.e. the post-edited translation.

Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

Segment ranking

In the WMT metrics task, human assessment based on segment ranking was often used. Human assessors were frequently asked to provide a complete ranking over all the candidate translations of the same source segment [33]. In the WMT13 shared tasks [20], five systems were randomised for the assessor to give a rank to their output. Each time the source segment and the reference translation were presented together with the candidate translations from the five systems. The assessors ranked the systems from 1 to 5, allowing tied scores. For each ranking, there was the potential to provide as many as 10 pairwise results if there were no ties. The collected pairwise rankings were then used to assign a corresponding score to each participating system to reflect the quality of the automatic translations. The assigned scores could also be used to reflect how frequently a system was judged to be better or worse than other systems when they were compared on the same source segment, according to the following formula:

$$\frac{\#\text{better pairwise ranking}}{\#\text{total pairwise comparison} - \#\text{ties comparisons}} \tag{2.1}$$

Han (2022) Thesis
'An investigation into multi-word expressions in machine translation'
https://doras.dcu.ie/26559

[91] noted that the lower agreements from WMT human assessment might be caused partially by the interval-level scales set up for the human assessor to make a quality judgement of each segment. For instance, the human assessor might be in a situation where neither of the two categories they were forced to choose is preferred. In light of this rationale, they proposed continuous measurement scales (CMS) for human TQA using fluency criteria. This was implemented by introducing the crowd-sourcing platform Amazon Mechanical Turk (MTurk), which has been popular in both NLP and multimedia research tasks [130], with some quality control methods such as the insertion of *bad-reference* and *ask-again*, and statistical significance testing. This methodology reportedly improved both intra-annotator and inter-annotator consistency. Detailed quality control methodologies, including statistical significance testing were documented in direct assessment (DA) [90, 92].

To achieve better human evaluation, moving from crowd-sourced DA to DA by professional linguists was implemented in some very recent shared tasks, e.g. WMT-QE2020 [254], with the expectation that this will produce more reliable reference scores and rankings, but they only implemented this for a small number of language pairs due to the high cost of the manual assessment.

Crowd-sourced Intelligence

- Fluency
- Quality Controls
- DA: crowd to experts

Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

In this thesis we focus on the challenge of MWEs in MT, since it has a broader connection and influence. For instance, it is connected to linguistic awareness, related to low-frequency phrases translation from statistical point of view, and as well as related to idiomaticity in translation.

Various definitions of MWEs have included both syntactic structure and semantic viewpoints from different researchers covering syntactic anomalies, non-compositionality, non-substitutability and ambiguity [53]. For instance, [11] **define MWEs** as "lexical items that: (i) can be decomposed into multiple lexemes; and (ii) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". MWEs have a broad coverage of linguistic phenomena, both syntax and semantics, such as compound nouns, named entities, verb particle constructions, discourse markers, collocations, lexical bundles, idioms and metaphors, and more. MWEs can appear in unexpected syntax and can lead to ambiguity. The syntactic richness and idiomaticity of MWEs also inspires MT and NLP researchers to design new methodologies, in addition to the challenges it presents [53].

Multi-word Expressions (MWEs)
- MT challenges
- HumanEval on MWEs

Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

Before moving to the first section, we provide a *working definition* of MWEs for our experimental investigation. Conventional researchers have defined MWEs from the perspectives of decompositionality and idiomaticity, as we

89

MWE
Working definition
- Multi-lingual setting

An Investigation into Multi-word Expressions in Machine Translation

mentioned in Section 1.4. Following this inspiration, from a multilingual setting, we define an MWE as *a group of lexemes that are formed into a pattern to carry one concept or expression.* This pattern can be semi-fixed or fixed, and it can be in a presentation of continuous lexemes or discontinuous ones.

Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

Chinese Examples of continuous MWEs

For instance, the Chinese MWE "洗衣機 (xǐ yī jī)" has three character components "洗 (xǐ)", "衣 (yī)", and "機 (jī)", of which "洗 (xǐ)" can be an independent word meaning "wash (something)", "衣 (yī)" is from the word '衣服 (yī fú)'" meaning "cloth(es)", and "機 (jī)" is from the word "機器 (jī qì)" meaning "machine(s)". All together, these three Chinese characters form a Chinese MWE carrying the concept of a "washing machine" which is a corresponding noun phrase (verb-ing + noun) terminology in English. In German, this is translated into one word "Waschmaschine" which belongs to one kind of MWEs that was listed in Section 1.4 called compound nouns. In Section 4.2,

MWEs as MT challenges

MWE Examples in English

- As examples of MWEs in English, there are (just to list a few)
- - verb-noun patterns ``jump the gun'' and ``kick the bucket'',
- - noun-noun combinations ``hit man'' and ``word salad'',
- - adjective-noun combinations like ``cold shoulder'' and ``hot dog'',
- - binomial expressions such as ``by and large'' and ``nuts and bolts'',

Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

Verbal MWEs

The working definition of vMWEs was introduced by PARSEME shared task on vMWE identification and discovery that has been organised almost each year since 2017 [180, 200]. It is defined as following: a vMWE has a verb as the head of the studied MWE term and functions as a verbal phrase, with examples such as “kick the bucket”. We also have more examples from our corpus creation some of which will be explained in Section 5.2.4 including “cutting capers” and “(beer) gone to his head”, etc.

AlphaMWE: A multi-lingual corpus with MWEs for MT testing

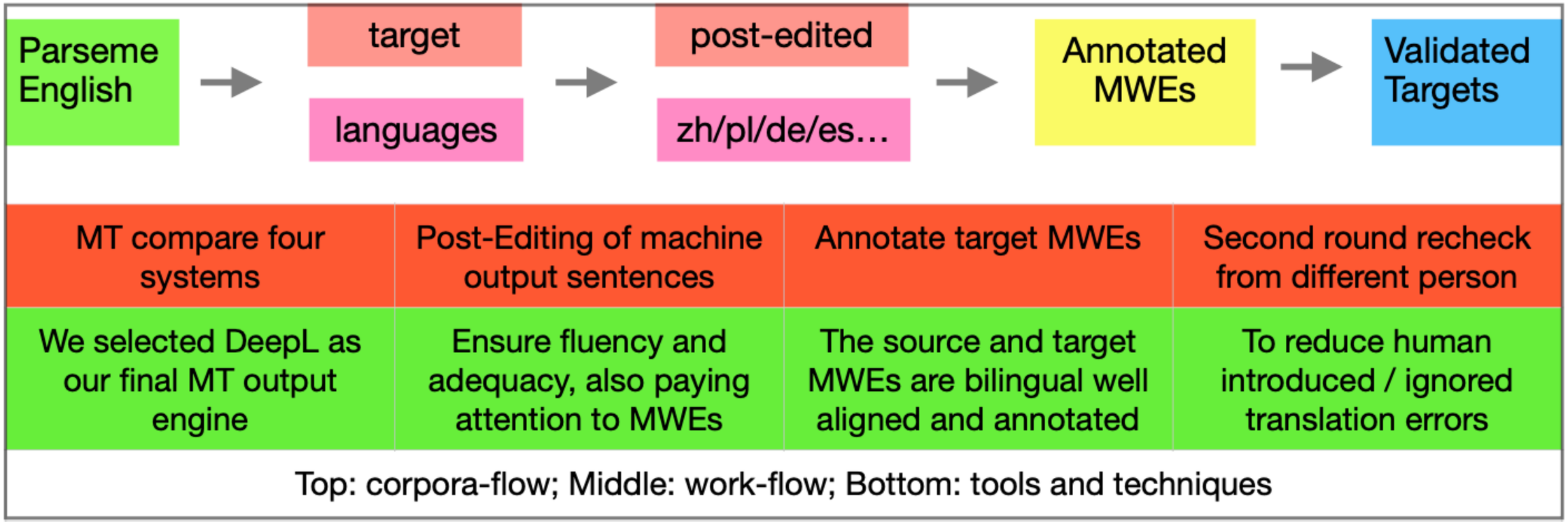


Figure 5.1: Workflow to prepare the AlphaMWE corpus.

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

Multi-word Expressions (MWEs) - HumanEval

Han, Smeaton, Jones (2020) AlphaMWE

- Error classifications on MT training MWEs related sentences
- Covering languages: English-Chinese/German/Polish
- English-Chinese: Common Sense, Super Sense, Abstract Phrases, Idiom, Metaphor, and Ambiguity.
- - Context-Unaware Ambiguity
- - Social/Literature-Unaware Ambiguity
- - Coherence-unaware Ambiguity (CohUA)

Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. AlphaMWE: Construction of Multilingual Parallel Corpora with MWE Annotations. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons*, pages 44–57, online. Association for Computational Linguistics.

HumanEval: MT on MWEs

Common sense

Source	At the corner of 72nd Street and Madison Avenue, he waved down a cab.
DeepL	在72街和麦迪逊大道的拐角处，他向一辆出租车招手。 Wave hand Zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù, tā xiàng yī liàng chū zū chē zhāo shǒu.
Bing	在72街和麦迪逊大道的拐角处，他挥手示意一辆出租车。 Wave hand to indicate zài 72 jiē hé mài dí xùn dà dào de guǎi jiǎo chù , tā huī shǒu shì yì yí liàng chū zū chē.
Google	在第72街和麦迪逊大街的拐角处，他挥舞着出租车。 Wave and dance with Zài dì 72 jiē hé mài dí xùn dà jiē de guǎi jiǎo chù, tā huī wǔ zhe chū zū chē.
Baidu	在72街和麦迪逊大街的拐角处，他挥手叫了一辆出租车。 Wave and got zài 72 jiē hé mài dí xùn dà jiē de guǎi jiǎo chù, tā huī shǒu jiào le yí liàng chū zū chē.

Han, Smeaton, Jones (2021) AlphaMWE https://aclanthology.org/2020.mwe-1.6

HumanEval: MT on MWEs

Super sense

Source	Each time he took a walk, he felt as though he were leaving himself behind, and by giving himself up to the movement of the streets, by reducing himself to a seeing eye, he was able to escape the obligation to think, and this, more than anything else, brought him a measure of peace, a salutatory emptiness within.
DeepL	每一次散步，他都觉得自己好像把自己抛在了脑后，把自己交给了街道的运动，把自己还原成了一只看得见的眼睛，他就能摆脱思考的义务，这比什么都重要，这给他带来了一定程度的平静，内心的空虚让人敬佩。(jiāo gěi; nèi xīn de kōng xū ràng rén jìng pèi)
Bing	每次散步时，他都觉得自己好像要把自己抛在脑后，通过向街道运动自在，通过把自己留眼，他能够逃避思考的义务，这比什么都重要，给他带来了一定程度的和平，一种有益的空虚。(zì zài; yǒu yì de kōng xū)
Google	每次散步时，他都会感觉自己好像在抛弃自己，投身于大街小巷，睁大了眼睛，这使他摆脱了思考的义务，而这， 最重要的是，他带给他一种和平的感觉，一种内在的称呼空虚。(tóu shēn yú; nèi zài de chēng hu kōng xū)
Baidu	每次他散步，他都觉得自己好像是在离开自己，把自己交给街道的流动，把自己变成一个有眼光的人，他就可以逃避思考的义务，而这比其他任何事情都能给他带来某种程度的平和，一种内在的致意的空虚。(nèi zài de zhì yì de kōng xū)
Ref.	每一次散步，他都觉得自己好像把自己抛在了脑后，投身于大街小巷，把自己还原成了一只看得见的眼睛，他就能摆脱思考的义务，这给他带来了某种程度的平静和内心悦纳的空无，远胜于其他。(nèi xīn yuè nà de kōng wú)

Gave to

Self-enjoy

devoted himself to

An inner calling emptiness

Gave to

devoted himself to

An inner emptiness you are happy with

Han, Smeaton, Jones (2021) AlphaMWE https://aclanthology.org/2020.mwe-1.6

HumanEval: MT on MWEs

Abstract phrase - needing of context

Source	Quinn had his doubts, but this was all he had to **go on**, his only bridge to the present.
DeepL	奎恩有他的疑虑，但这是他唯一的依靠，是他通往现在的唯一桥梁。(wéi yī de yī kào) Only relying thing
Bing	奎因有他的怀疑，但这就是他必须**去**的，他唯一的桥梁到现在。 (bì xū qù de) Has to go
Google	奎因有他的疑惑，但这就是他所要**做**的，是他通往现在的唯一桥梁。 (suǒ yào zuò de) Need to do
Baidu	奎因有他的疑虑，但这是他所要**做**的，是他通往现在的唯一桥梁。 (suǒ yào zuò de) Need to do
Ref.	奎恩曾有他的疑虑，但这是他**开展工作**的所有依据，是他通往现在的唯一桥梁。(**kāi zhǎn gōng zuò** de suǒ yǒu yī jù) All he had to carry on the work / to work from (finding someone)

Sentence level MT would make no much sense in this situation.
- context aware MT, and document-level MT

Han, Smeaton, Jones (2021) AlphaMWE https://aclanthology.org/2020.mwe-1.6

Idiom - examples

Source	I was smoking my pipe quietly by my dismantled steamer, and saw them all cutting capers in the light, with their arms lifted high, when the stout man with mustaches came tearing down to the river, a tin pail in his hand, assured me that everybody was 'behaving splendidly, splendidly, dipped about a quart of water and tore back again.
DeepL	我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下高举双臂切着荠菜，这时，那个留着胡子的壮汉撕扯着来到河边，手里拿着一个铁皮桶，向我保证说，大家都'表现得很精彩，很精彩，浸了大约一夸脱的水，又撕扯回来了。 (qiē zhe jì cài; sī chě huí lái) Cutting 'Shepherd's Purse'
Bing	我悄悄地用我拆除的蒸笼抽我的烟斗，看到他们所有的切割披肩在光中，他们的手臂抬起来高，当粗壮的人用胡子来到河中，一个锡桶在他的手中，向我保证，每个人都'表现辉煌，辉煌，浸约夸脱的水，并再次撕毁。(qiē gē pī jiān; zài cì sī huǐ) Cutting shawl
Google	我正用拆卸的蒸锅安静地抽着烟斗，看到他们都在灯光下割下刺山柑，双臂高高举起，当时那留着小胡子的矮壮男人向河边撕下，手里拿着一个铁桶，向我保证 每个人都表现得非常出色，表现得非常出色，蘸了一夸脱的水，然后又撕了回来。(gē xià cì shān gān; sī le huí lái) Cutting 'capers'
Baidu	我正静静地在我拆掉的汽船旁抽着烟斗，看见他们都高举着胳膊，在灯光下割着山柑，这时那个长着胡子的胖男人手里拿着一个锡桶，朝河里跑来，向我保证每个人都表现得很好，很漂亮，蘸了一夸脱水，然后又往回跑。(gē zhe shān gān; wǎng huí pǎo) Cutting 'capers'
Ref.	我在拆开的汽船旁静静地抽着烟斗，看到他们都在灯光下欢呼雀跃，高举双臂，这时，那个留着胡子的大块头，手里拿着一个铁皮桶，快速来到河边，向我确保大家都"表现得很精彩，很精彩"，他浸了大约一夸脱的水，又快速回去了。(huān hū què yuè; kuài sù huí qù)

cut a caper: make a playful skipping movement

Han, Smeaton, Jones (2021) AlphaMWE

HumanEval: MT on MWEs

Metaphor examples

Source	The what? Auster laughed, and in that laugh everything was suddenly blown to bits. The chair was comfortable, and the beer had gone slightly to his head.
DeepL	那个什么？奥斯特笑了，在这笑声中，一切突然被炸得粉碎。(bèi zhà dé fěn suì) 椅子很舒服，啤酒已经微微到了他的头上。 (wēi wēi dào le tā de tóu shàng)
Bing	什么？奥斯特笑了，在笑，一切都突然被吹成位。(bèi chuī chéng wèi) 椅子很舒服，啤酒稍微到他的头去了。 (shāo wēi dào tā de tóu qù le)
Google	什么啊 Auster笑了起来，在那笑声中，一切突然被炸碎了。 (bèi zhà suì le) 椅子很舒服，啤酒微微飘到他的头上。(wēi wēi piāo dào tā de tóu shàng)
Baidu	什么？奥斯特笑了，在那笑声中，一切都突然被炸成碎片。(bèi zhà chéng suì piàn) 椅子很舒服，啤酒已经稍稍流到他的头上了。(shāo shāo liú dào tā de tóu shàng le)
Ref.	那个什么？奥斯特笑了，在这笑声中，一切突然化为乌有。(huà wéi wū yǒu) 椅子很舒服，啤酒已经微微让他上了头。 (wēi wēi ràng tā shàng le tóu)

Han, Smeaton, Jones (2021) AlphaMWE https://aclanthology.org/2020.mwe-1.6

Category-VI: Context-Unaware Ambiguity (CUA)

In this case, the context, i.e. the background information, is needed for correct translation of the sentence. For instance, see Figure 5.9. DeepL gives the translation "it did not give me time though", while Bing and GoogleMT give the same translation "it/this did not give me one day's time" and Baidu outputs a grammatically incorrect sentence. From the pre-context, we understand that it means the speaker "did not feel that is special to him" or "did not have affection of that" after *all the Mormon missionary's effort towards him*. Interestingly, there is a popular Chinese idiom (slang) that matches this meaning very well "不是我的菜 (bù shì wǒ de cài, literally *not my dish*)". From this point of view, the context based MT model deserves some more attention, instead of only focusing on the sentence level. When we tried to put all background context information as shown in Figure 5.9 into the four MT models, they produce as the same output for this studied sentence, as for sentence level MT. This indicates that current MT models still focus on sentence-by-sentence translation when meeting paragraphs, instead of using context inference.

HumanEval: MT on MWEs

Han, Smeaton, Jones (2021) AlphaMWE
https://aclanthology.org/2020.mwe-1.6

Context-Unaware Ambiguity (CUA)

Source	But it did not give me the time of day.
DeepL	但它并没有给我时间。 (gěi wǒ shí jiān)
Bing	但它没有给我一天的时间。 (gěi wǒ yī tiān de shí jiān)
Google	但这没有给我一天的时间。 (gěi wǒ yī tiān de shí jiān)
Baidu	但它没有给我一天中的时间。 (gěi wǒ yī tiān zhōng de shí jiān)
Ref.	但我没有感到这个对于我特殊 / 但这不是我的菜。(gǎn dào zhè ge duì yú wǒ tè shū / … wǒ de cài)
Context	An old Mormon missionary in Nauvoo once gripped my knee hard as we sat side by side, and he put his arm about me and called me "Brother." We'd only met ten minutes before. He took me to his good bosom. His eyes began to mist. I was a prospect, an exotic prospect in old tennis shoes and a sweatshirt. His heart opened to me. It opened like a cuckoo clock. But it did not …

Figure 5.9: MT issues with MWEs: context-unaware ambiguity

Han, Smeaton, Jones (2021) AlphaMWE https://aclanthology.org/2020.mwe-1.6

Social/Literature-Unaware Ambiguity (SLUA), Domain Knowledge

Category-VI: Social/Literature-Unaware Ambiguity (SLUA)

In this case, social knowledge of current affairs from news, or literature knowledge about some newly invented entities / phrases is required in order to get correct translation output. For instance, Figure 5.10 includes two sentences, one from politics and another from literature.

In the first sentence, "de-gnoming" is a word from Harry Potter, invented

---

Han, Smeaton, Jones (2021) AlphaMWE https://aclanthology.org/2020.mwe-1.6

Social/Literature-Unaware Ambiguity (SLUA), Domain Knowledge

Source	The moment they know the de-gnoming's going on they storm up to have a look. Then someone says that it can't be long now before the Russians write Arafat off.
DeepL	他们一知道去核的事，就会冲上去看一看。 (qù hé) 然后有人说，现在用不了多久，俄罗斯人就会**把**阿拉法特**注销**。(**bǎ** ā lā fǎ tè **zhù xiāo**)
Bing	当他们知道去诺格明是怎么回事， 他们冲了起来看看。 (qù nuò gé míng) 然后有人说，现在俄罗斯人要不长了，就**把**阿拉法特**注销**了。(**bǎ** ā lā fǎ tè **zhù xiāo** le)
Google	当他们知道正在逐渐消失的那一刻，他们便冲上去看看。 (zhèng zài zhú jiàn xiāo shī) 然后有人说，不久之后俄罗斯人**将**阿拉法特**注销**。 (**jiāng** ā lā fǎ tè **zhù xiāo**)
Baidu	他们一知道德格诺明正在进行，就冲上去看一看。 (dé gé nuò míng) 然后有人说，俄国人很快就会把阿拉法特**一笔勾销**了。 (bǎ ā lā fǎ tè **yī bǐ gōu xiāo** le)
Ref.	一知道去地精的事在进行，他们就冲上去观看。 (qù dì jīng) 然后有人说，现在用不了多久，俄罗斯人就会**把**阿拉法特**下課** / 让…下台。 (**bǎ** ā lā fǎ tè **xià kè;** ràng…xià tái)

Figure 5.10: MT issues with MWEs: social/literature-unaware ambiguity

Han, Smeaton, Jones (2021) AlphaMWE https://aclanthology.org/2020.mwe-1.6

Social/Literature-Unaware Ambiguity (SLUA), Domain Knowledge

by its author, to refer to the process of ridding a garden of gnomes, a small magical beast. Without this literary knowledge it is not possible to translate the sentence correctly. For instance, even though this sentence is from a very popular novel that has been translated into many languages, DeepL translated it as "去核 (qù hé, de-nuclear)", Bing translated it as "去诺格明 (qù nuò gé míng, *de*-nuògémíng" where "nuògémíng" is a simulation of the pronunciation of "gnoming" in a Chinese way, Baidu translated it as "德格诺明 (dé gé nuò míng)" which is the simulation of the pronunciation of the overall term "de-gnoming".

Han, Smeaton, Jones (2021) AlphaMWE https://aclanthology.org/2020.mwe-1.6

MT on MWEs

Category-VI: Coherence-unaware Ambiguity (CohUA)

Source	Two months ago I had to have an operation for a serious ***complaint.***
DeepL	两个月前，我因为一次严重的**投诉**不得不做手术。(**tóu sù** … zuò shǒu shù)
Bing	两个月前，我不得不做一个严重的**投诉**手术。 (zuò … **tóu sù** shǒu shù)
Google	两个月前，我不得不接受一次手术以应对严重的**投诉**。 (jiē shòu yī cì shǒu shù … **tóu sù**)
Baidu	两个月前，我因为严重的**投诉**不得不动手术。 (**tóu sù** … dòng shǒu shù)
Ref.	两个月前，我因为一次严重的**症状**不得不做手术。(**zhèng zhuàng** … zuò shǒu shù)

Coherence of the sentence, natural language inference will help the disambiguation.

This kind of MWE ambiguity can be solved by the coherence of the sentence itself, for instance, the example in Figure 5.11. The four MT models all translated the vMWE itself "have an operation" correctly in meaning preservation by "做/接受/动手术 (zuò/jiē shòu/dòng shǒu shù)" just with different Chinese word choices. However, none of the MT models translated the "reason of the operation", i.e., "complaint" correctly. The word complaint has two most commonly used meanings "a statement that something is unsatisfactory or unacceptable" or "an illness or medical condition" and all four models chose the first one. According to simple logic of social life, people do not need to "have an operation" due to "a statement", instead their "medical condition" should have been chosen to translate the word "complaint". Because of the in-

Multidimensional Quality Metrics (MQM) framework

Table of Contents

Too large and too many dimensions, how about good enough?
- a hint to '**HOPE**' metric https://github.com/lHan87/HOPE

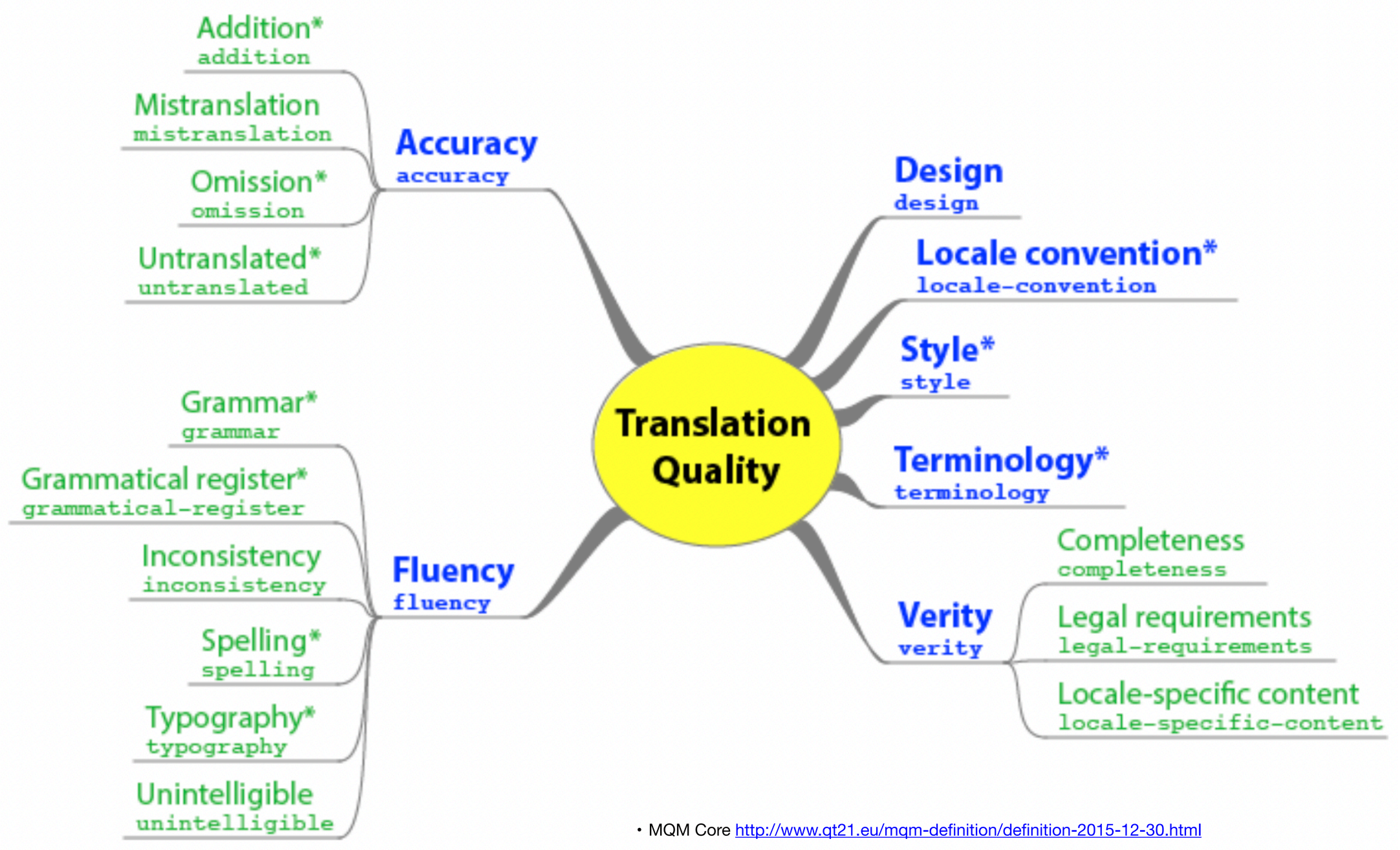
Editors

- Arle Lommel (DFKI)
- Aljoscha Burchardt (DFKI)
- Hans Uszkoreit (DFKI)

Contributors

- Kim Harris (text&form)
- Alan K. Melby (LTAC Global)
- Attila Görög (TAUS)
- Serge Gladkoff (Logrus)
- Leonid Glazychev (Logrus)

- http://www.qt21.eu/mqm-definition/definition-2015-12-30.html



- MQM Core http://www.qt21.eu/mqm-definition/definition-2015-12-30.html
- Now MQM 2.0 https://themqm.org

The 19 issues are defined in the MQM core as follows:

- *Accuracy (`accuracy`)*
  - *Addition (`addition`)*
  - *Mistranslation (`mistranslation`)*
  - *Omission (`omission`)*
  - *Untranslated (`untranslated`)*
- *Design (`design`)*
- *Fluency (`fluency`)*
  - *Grammar (`grammar`)*
  - *Grammatical register (`grammatical-register`)*
  - *Inconsistency (`inconsistency`)*
  - *Spelling (`spelling`)*
  - *Typography (`typography`)*
  - *Unintelligible (`unintelligible`)*
- *Locale convention (`locale-convention`)*
- *Style (`style`)*
- *Terminology (`terminology`)*
- *Verity (`verity`)*
  - *Completeness (`completeness`)*
  - *Legal requirements (`legal-requirements`)*
  - *Locale-specific content (`locale-specific-content`)*

- http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

MQM vs TAUS DQF

10. TAUS DQF subset (non-normative)

The TAUS DQF Error Typology is a recognized subset of MQM, developed and maintained by the Translation Automation User Society (TAUS) based on input from its members.

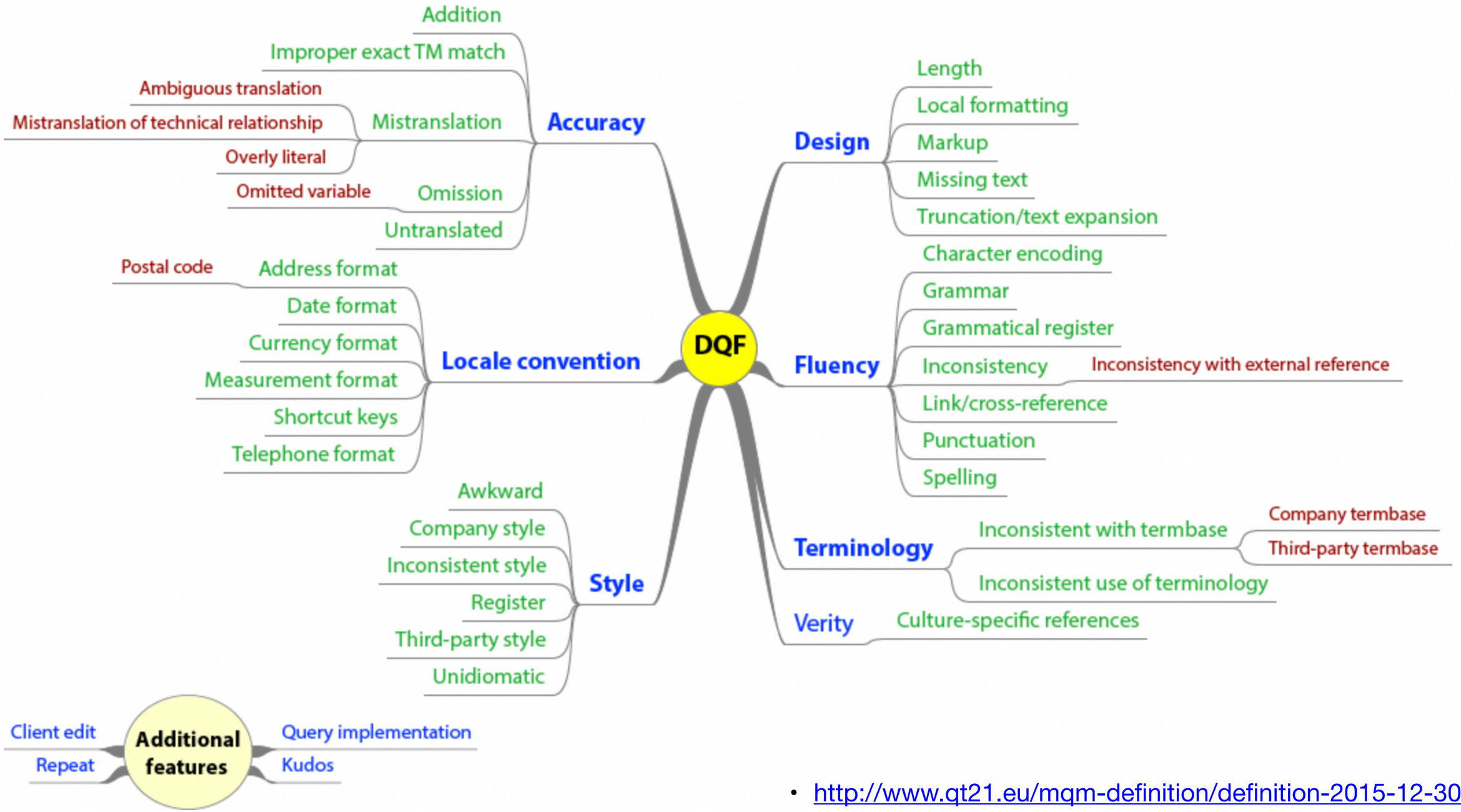
Previous versions of TAUS DQF and MQM were not compatible. As of revision 0.9, compatibility between the two has been achieved. The harmonization process required substantial modification to both MQM and DQF, but now DQF, with the exception of the “kudos” feature noted below, is a fully conformant subset of MQM.

The DQF tools check six issue types. If only these issue types are used, they correspond directly to MQM dimensions, as follows:

- DQF *Accuracy* (formerly *Adequacy*) = MQM *Accuracy* (`accuracy`)
- DQF *Design* (formerly *Layout*) = MQM *Design* (`design`)
- DQF *Fluency* (formerly *Language*) = MQM *Fluency* (`fluency`)
- DQF *Locale convention* (formerly *Country standards*) = MQM *Locale convention* (`locale-convention`)
- DQF *Style* = MQM *Style* (`style`)
- DQF *Terminology* = MQM *Terminology* (`terminology`)

- http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

MQM also supports additional levels of issues, as shown in the following graphic:



- http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

MQM 2.0 = ASTM WK46396

Analytic Evaluation of Translation Quality

The work item: www.astm.org/workitem-wk46396

More detailed information: https://themqm.org/

Task-oriented HOPE metric

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

Serge Gladkoff[1], **Lifeng Han**[2]
[1] Logrus Global LLC
[2] ADAPT Research Centre, Dublin City University
serge.gladkoff@logrusglobal.com & lifeng.han3@mail.dcu.ie

quality of worse than premium quality translations. In this work, we introduce **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 only a limited number of commonly occurring error types, and use a scoring model with geometric progression of error penalty points (EPPs) reflecting error severity level to each translation unit. The initial experimental work carried out on English-Russian language pair MT outputs on marketing content type of text from highly technical domain reveals that our evaluation framework is quite effective in reflecting the MT output quality regarding both overall system-level performance and segment-level transparency, and it increases the IRR for error type interpretation. The approach has several key advantages, such as 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, low-cost and faster application, as well as higher IRR. Our experimental data is available at `https://github.com/lHan87/HOPE`.

HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation S Gladkoff, L Han - arXiv preprint arXiv:2112.13833, 2021

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.

$$\text{EPPTU} = \sum_i Error_i \times Severity(i) \quad (4)$$

where $Severity(i)$ is the severity level of $Error_i$. Each TU has its own EPPTU not depending on other TUs. Importantly, repeated errors in different TUs are not counted as one error, because MT outputs experience stochastic behavior and errors are not made consistently. One and the same error may repeat itself, but more often is mixed with other instances of a similar error. The system-level score of HOPE is calculated by the sum of overall segment-level EPPTUs:

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

3.3.1. Segment/Sentence-Level HOPE

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

3.3.2. Word-Level HOPE

In addition to the segment-level HOPE deployment, we also design a word level HOPE evaluation in its application. 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 them three categories. Both segment/sentence-level

HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation S Gladkoff, L Han - arXiv preprint arXiv:2112.13833, 2021

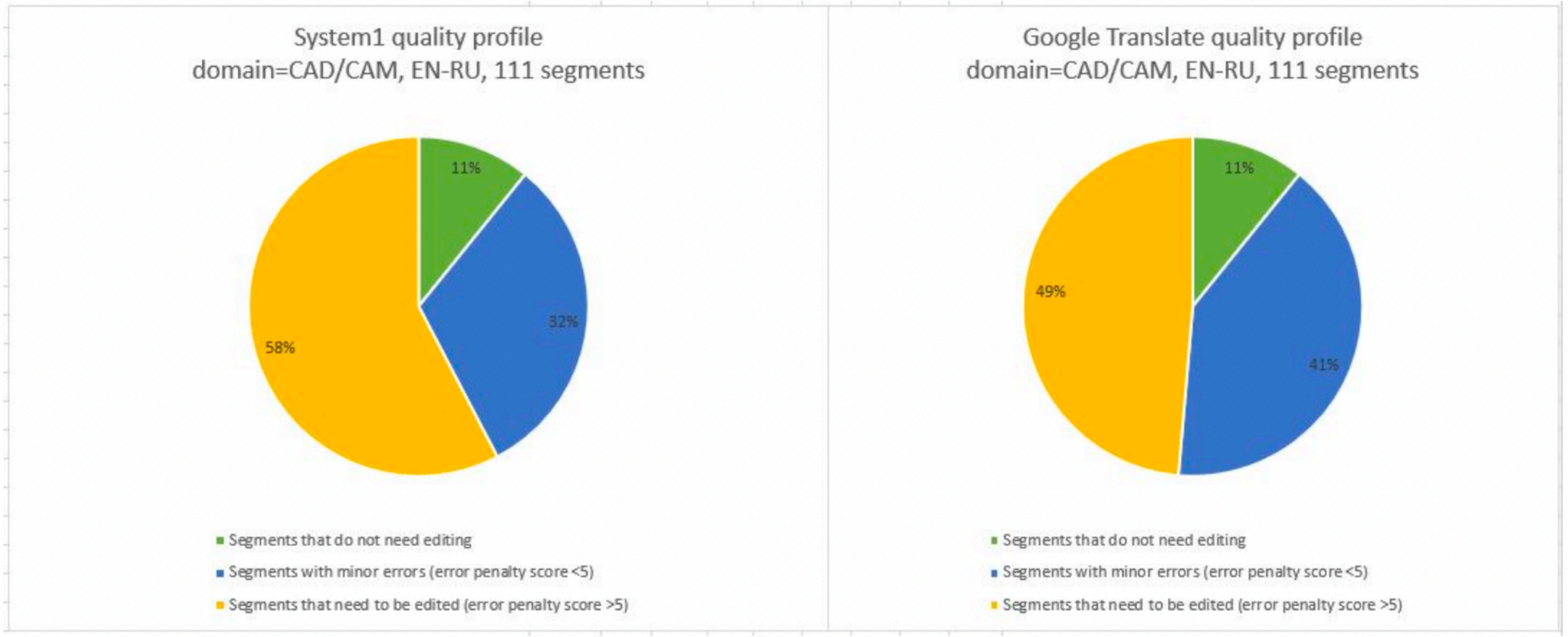


Figure 3: Task-I: Comparison of System1 and Google Translate HOPE Quality Profiles.

HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation S Gladkoff, L Han - arXiv preprint arXiv:2112.13833, 2021

	Unchanged	Minor	Major	
Total	278	275	118	
%	41%	41%	18%	
Wordcount	691	1718	930	3339
	21%	51%	28%	

Figure 4: Quality Indicators via Segment vs Word Level HOPE: number counts and percentage.

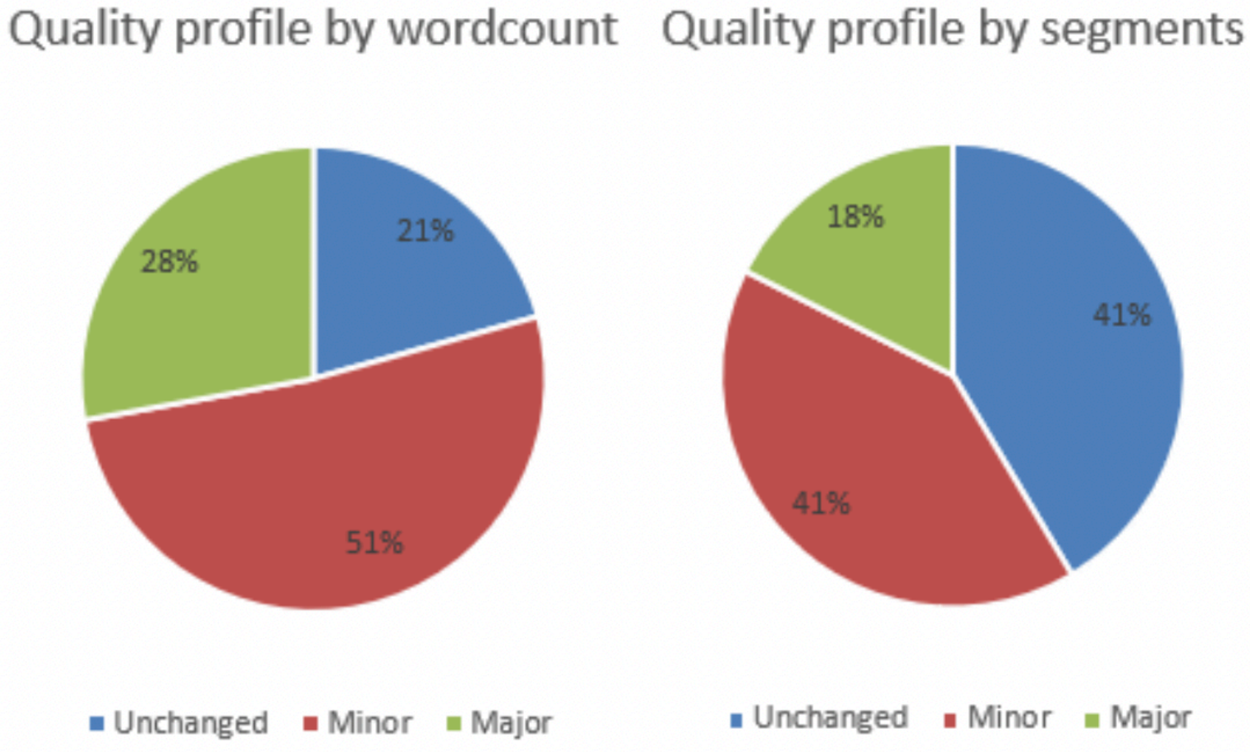
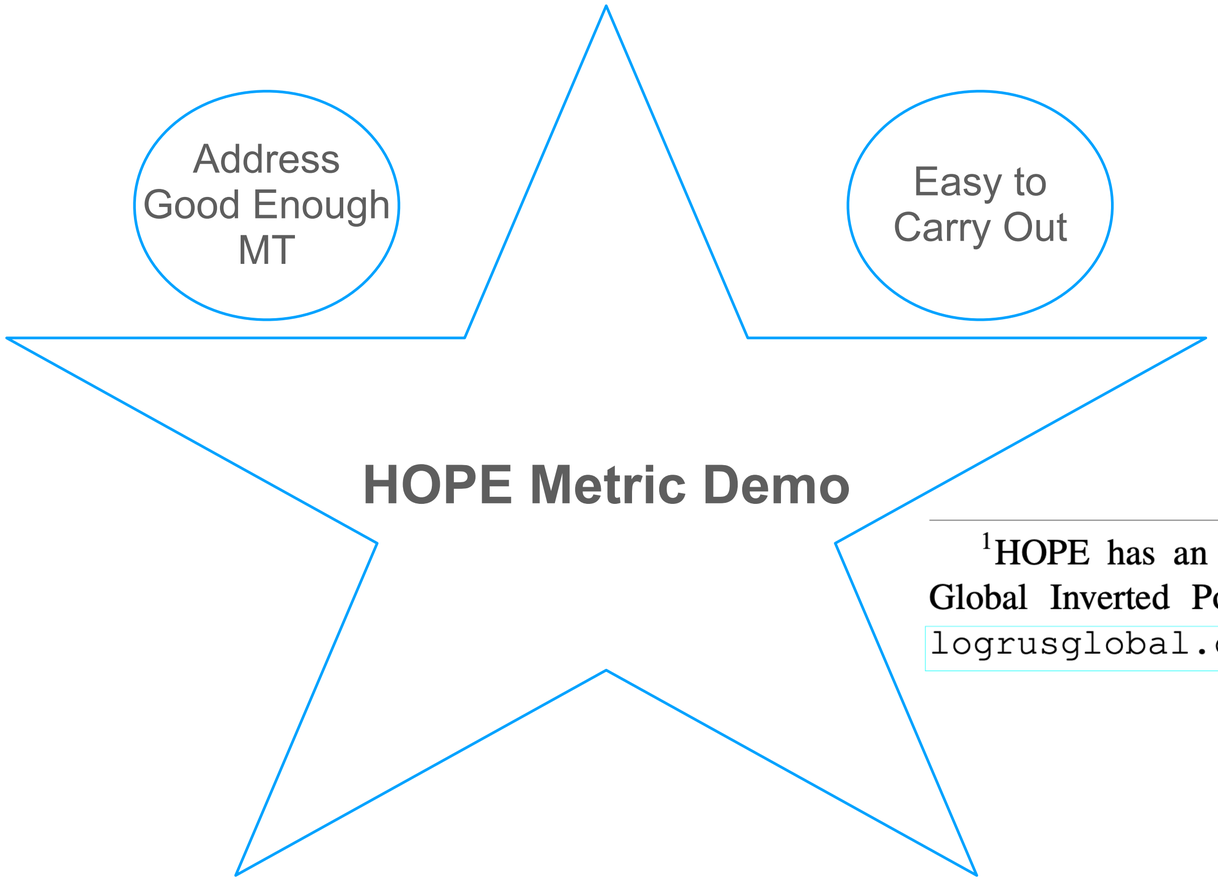


Figure 5: Quality Indicators via Segment vs Word Level HOPE: percentage.

HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation S Gladkoff, L Han - arXiv preprint arXiv:2112.13833, 2021



[1]HOPE has an alternative name: LOGIPEM (LOgrus Global Inverted Post-Editing Metrics) *ref.* `https://logrusglobal.com/`

S Gladkoff, L Han (2022) HOPE: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation. LREC22. https://arxiv.org/abs/2112.13833
Data: https://github.com/lHan87/HOPE

Document-level HumEval

Castilho (2021)

- Issues on IAA (inter-annotator-agreement)
- - using adequacy, fluency, error mark-up (binary, type), pairwise ranking (GoogleMT, DeepL MT)
- - using post-task questionnaire: 10 statements on a scale [1, 6]
- Misevaluation “when sentences are evaluated out of context”

Castilho (2021) “Towards Document-Level Human MT Evaluation: On the Issues of Annotator Agreement, Effort and Misevaluation” https://aclanthology.org/2021.humeval-1.4.pdf

Another work on test-set: Maja Popović and Sheila Castilho. 2019. Challenge Test Sets for MT Evaluation. In *Proceedings of Machine Translation Summit XVII: Tutorial Abstracts*

Castilho (2021): questionnaires

- 1. Understanding the meaning of the source [in the random sentences/in each sentence, with access to the full document] in general was
- 2. Understanding the meaning of the translated [in the random sentences/in each sentence, with access to the full document] in general was
- 3. Recognising the adequacy problems [in the random sentences/in each sentence, with access to the full document] in general was
- 4. Recognising fluency problems [in the random sentences/in each sentence, with access to the full document] in general was
- 5. Spotting errors [in the random sentences/in each sentence, with access to the full document] in general was
- 6. Choosing the best of two translations [in the random sentences/in each sentence, with access to the full document] was
- 7. In general, assessing the translation quality on a [sentence/document] level was (difficulty)
- 8. For me, assessing the translation quality on a [sentence/document] level was (fatigue)
- 9. I was confident with every assessment I provided for the [sentence/document] level evaluation tasks
- 10. I could have done a more accurate assessment if I [had had access to the full text/was assessing random sentences]

There has been some criticism of the traditional human TQA methods because they fail to reflect real problems in translation by assigning scores and ranking several candidates from the same source [222]. Instead, [222] designed a new methodology by asking human assessors to mark all problematic parts of candidate translations, either words, phrases, or sentences. Two questions that were typically asked of the assessors related to *comprehensibility* and *adequacy*. The first criterion considers whether the translation is understandable, or understandable but with errors; the second criterion measures if the candidate translation has different meaning to the original text, or maintains the meaning

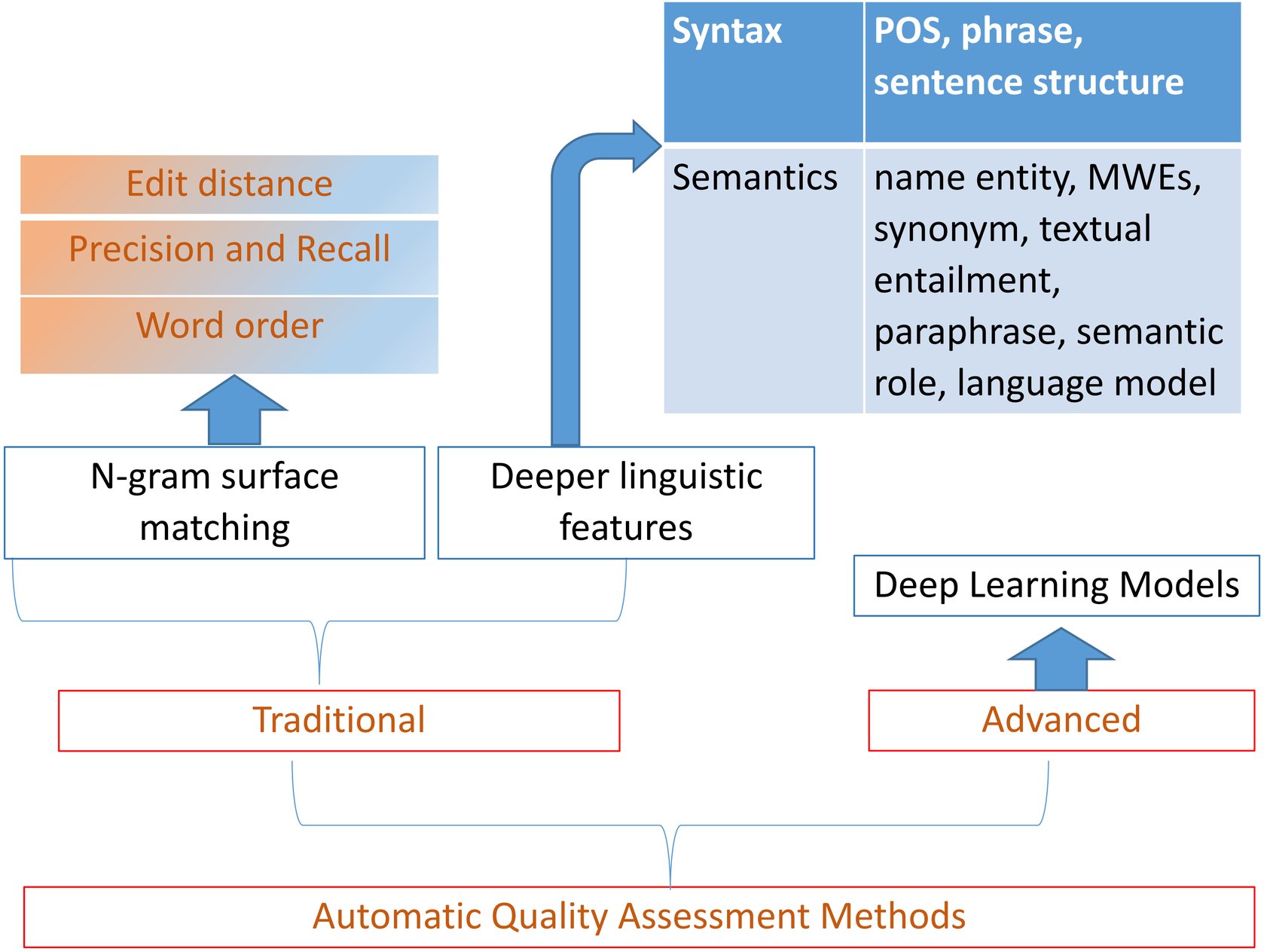
Revisiting traditional criteria

[222] Maja Popović. "Informative Manual Evaluation of Machine Translation Output". In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, Dec. 2020, pp. 5059–5069. DOI: 10.18653/v1/2020.coling-main.444.

[223] Maja Popović. "Relations between comprehensibility and adequacy errors in machine translation output". In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, Nov. 2020, pp. 256–264. DOI: 10.18653/v1/2020.conll-1.19.

but with errors. Both criteria take into account whether parts of the original text are missing in translation. Under a similar experimental setup, [223] also summarised the most frequent error types that the annotators recognised as misleading translations.

Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559



Ref: Han (2022) Thesis 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559

Automatic TQA/metrics

- n-gram string match: carried out on word surface level
    - editing distance (TER, PER, HTER)
    - precision, recall, f-score, word order (BLEU, METEOR, chrF)
    - augmented/hybrid (LEPOR)

- linguistic features: to capture syntax and semantics
    - syntax: POS, chunk, sentence structure (hLEPOR)
    - semantics: dictionary, paraphrase, NE, MWE, entailment, language model, semantic roles (MEANT)

- deep learning / ML models
    - LSTM, NNs, RNNs, BERTs: include syntactic and semantic features
    - Model distillation

AutoEval

N-gram string matches and synonyms usage

The widely used evaluation metric BLEU (Papineni et al., 2002) is based on the degree of n-gram overlapping between the strings of words produced by the machine and the human translation references at the corpus level. BLEU computes the precision for n-gram of size 1-to-4 with the coefficient of brevity penalty (BP).

$$\text{BLEU} = \text{BP} \times \exp \sum_{n=1}^{N} \lambda_n \log \text{Precision}_n, \quad (8)$$

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ e^{1-\frac{r}{c}} & \text{if } c <= r. \end{cases} \quad (9)$$

words that are synonyms of each other. To measure how well-ordered the matched words in the candidate translation are in relation to the human reference, METEOR introduces a penalty coefficient by employing the number of matched chunks.

$$\text{Penalty} = 0.5 \times \left(\frac{\#\text{chunks}}{\#\text{matched unigrams}}\right)^3, \quad (12)$$

$$\text{MEREOR} = \frac{10PR}{R+9P} \times (1 - \text{Penalty}). \quad (13)$$

(Banerjee and Lavie, 2005)

AutoEval

N-gram string matches

(Turian et al., 2006) conducted experiments to examine how standard measures such as precision and recall and F-measure can be applied for evaluation of MT and showed the comparisons of these standard measures with some existing alternative evaluation measures. F-measure is the combination of precision (P) and recall (R), which is firstly employed in the information retrieval and latterly has been adopted by the information extraction, MT evaluation and other tasks.

$$F_\beta = (1 + \beta^2)\frac{PR}{R + \beta^2 P} \qquad (11)$$

Automatic TQA/metrics - our work

N-gram string match: augmented/hybrid

LEPOR, nLEPOR: **Le**ngth Penalty, **P**recision, n-gram **Po**sition difference Penalty and **R**ecall. (n-gram P and R: nLEPOR) https://github.com/poethan/LEPOR/

Deeper linguistic features:

hLEPOR: word level + PoS <Harmonic mean of factors>

HPPR, word level + phrase structure

Using DL based model distillation:

cushLEPOR: customising hLEPOR to HumanEval or pre-trained LMs.

News: the python version implementation released this year in : https://pypi.org/project/hLepor/

Han et al. "LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors" in COLING2012.
hLEPOR: Language-independent Model for Machine Translation Evaluation with Reinforced Factors, in MT summit 2013
HPPR: Phrase Tagset Mapping for French and English Treebanks and Its Application in Machine Translation Evaluation. In GSCL2013
cushLEPOR: customising hLEPOR metric using Optuna for higher agreement with human judgments or pre-trained language model LaBSE. WMT2022

Example of word surface MTE

$$LEPOR = LP \times NPosPenal \times Harmonic(\alpha R, \beta P)$$

$$LP = \begin{cases} e^{1-\frac{r}{c}} & if \ \ c < r \\ 1 & if \ \ c = r \\ e^{1-\frac{c}{r}} & if \ \ c > r \end{cases}$$

- LP: length penalty
- NPosPenal: n-gram based position difference penalty
- R: recall
- P: precision
- LEPOR(A): system-level A
- LEPOR(B): system-level B

$$NPosPenal = e^{-NPD}$$

$$NPD = \frac{1}{Length_{output}} \sum_{i=1}^{Length_{output}} |PD_i|$$

$$Harmonic(\alpha R, \beta P) = (\alpha + \beta)/(\frac{\alpha}{R} + \frac{\beta}{P})$$

$$P = \frac{common\_num}{system\_length}$$

$$R = \frac{common\_num}{reference\_length}$$

$$\overline{LEPOR_A} = \frac{1}{SentNum} \sum_{i=1}^{SentNum} LEPOR_i$$

$$\overline{LEPOR_B} = \overline{LP} \times \overline{PosPenalty} \times \overline{Harmonic(\alpha R, \beta P)}$$

Han et al. (2012) "LEPOR: A Robust Evaluation Metric for Machine Translation with Augmented Factors" in COLING. https://aclanthology.org/C12-2044.pdf Github: https://github.com/poethan/LEPOR

Output Sentence: $W = \{w_1 w_2 w_3 \dots w_{m_1} \mid m_1 \in (1, \infty)\}$
Reference Sentence: $W^r = \{w_1^r w_2^r w_3^r \dots w_{m_2}^r \mid m_2 \in (1, \infty)\}$
$\forall x \in (1, \infty)$, The Alignment of word w_x:

---

if $\forall y \in (1, \infty): w_x \neq w_y^r$ // ∀ *means for each,* ∃ *means there is/are*
    $(w_x \rightarrow \emptyset)$; // → *shows the alignment*
elseif $\exists! y \in (1, \infty): w_x = w_y^r$ // ∃! *means there exists exactly one*
    $(w_x \rightarrow w_y^r)$;
elseif $\exists y_1, y_2 \in (1, \infty): (w_x = w_{y_1}^r) \wedge (w_x = w_{y_2}^r)$ // ∧ *is logical conjunction, and*
    foreach $k \in (-n, -1) \cup (1, n)$
        foreach $j \in (-n, -1) \cup (1, n)$
            if $\exists k_1, k_2, j_1, j_2: (w_{x+k_1} = w_{y_1+j_1}^r) \wedge (w_{x+k_2} = w_{y_2+j_2}^r)$
                if $Distance(w_x, w_{y_1}^r) \leq Distance(w_x, w_{y_2}^r)$
                    $(w_x \rightarrow w_{y_1}^r)$;
                else
                    $(w_x \rightarrow w_{y_2}^r)$;
            elseif $\exists k_1, j_1: (w_{x+k_1} = w_{y_1+j_1}^r) \wedge (\forall k_2, j_2: (w_{x+k_2} \neq w_{y_2+j_2}^r))$
                $(w_x \rightarrow w_{y_1}^r)$;
            else // *i.e.* $\forall k_1, k_2, j_1, j_2: (w_{x+k_1} \neq w_{y_1+j_1}^r) \wedge (w_{x+k_2} \neq w_{y_2+j_2}^r)$
                if $Distance(w_x, w_{y_1}^r) \leq Distance(w_x, w_{y_2}^r)$
                    $(w_x \rightarrow w_{y_1}^r)$;
                else
                    $(w_x \rightarrow w_{y_2}^r)$;
else // *when more than two candidates, the selection steps are similar as above*

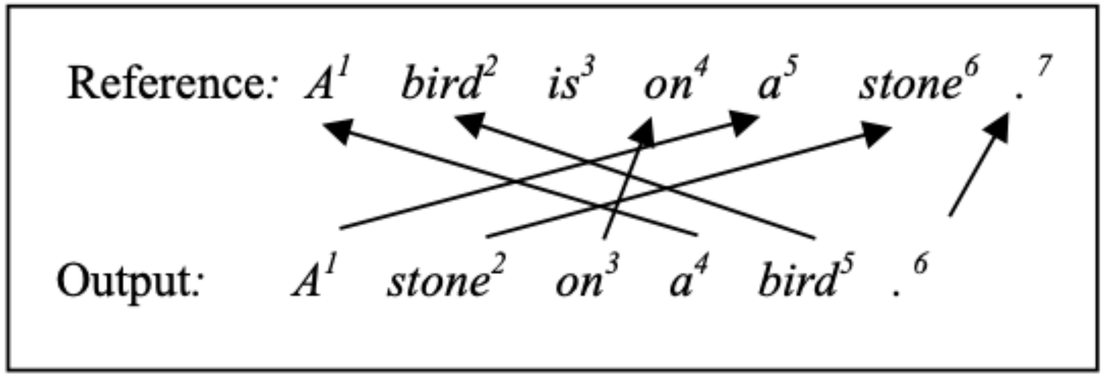
N-gram based word alignment
Considering context mapping

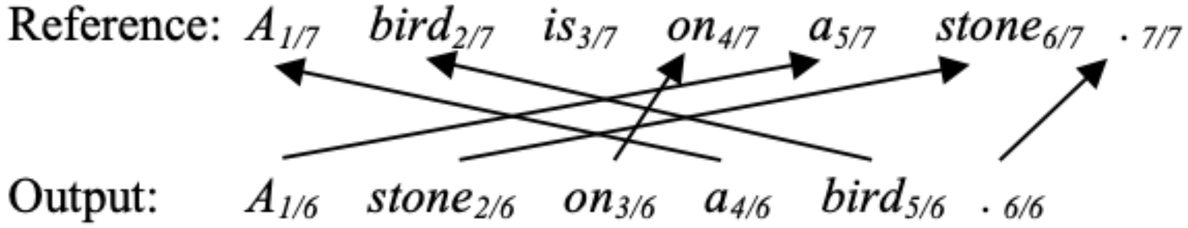
$for\ x = 1 \cdots l_h$
  $for\ y = 1 \cdots l_s$
    $if\ unmatch(tag_x, tag_y^s)$
      $Align(tag_x, \emptyset)$;
    $elseif\ match(tag_x, tag_y^s)$
      $Align(tag_x, tag_y^s)$;
    $else$ // more than one candidates matching tag_x $(i \geq 2)$
      $\boldsymbol{LookforFinalAlign}(tag_x, < tag_{y_1}^s, \dots, tag_{y_i}^s >)$
  $endfor$
$endfor$

$\boldsymbol{LookforFinalAlign}(tag_x, < tag_{y_1}^s, \dots, tag_{y_i}^s >)$
$for\ p = -N_1, \dots, -1, 1, \cdots, N_1$
  $for\ q = -N_1, \dots, -1, 1, \cdots, N_1$
    $if\ unmatch(tag_{x+p}, < tag_{y_1+q}^s, \dots, tag_{y_i+q}^s >)$ //no candidate tag has neighbor match
      $Align(tag_x, tag_{y_j}^s)\ when\ y_j\ (1 < j < i)\ has\ the\ shortest\ distance\ with\ x$;
    $elseif\ match(tag_{x+p}, tag_{y_j+q}^s)$ //only $tag_{y_j}^s$ has neighbor match with tag_x
      $Align(tag_x, tag_{y_j}^s)$;
    $else$ // more than two candidates have neighbor match with tag_x
      $Align(tag_x, tag_{y_j}^s)\ when\ y_j (1 < j < i)\ has\ the\ shortest\ distance\ with\ x$;
  $endfor$
$endfor$

N-gram based word alignment
Considering context mapping

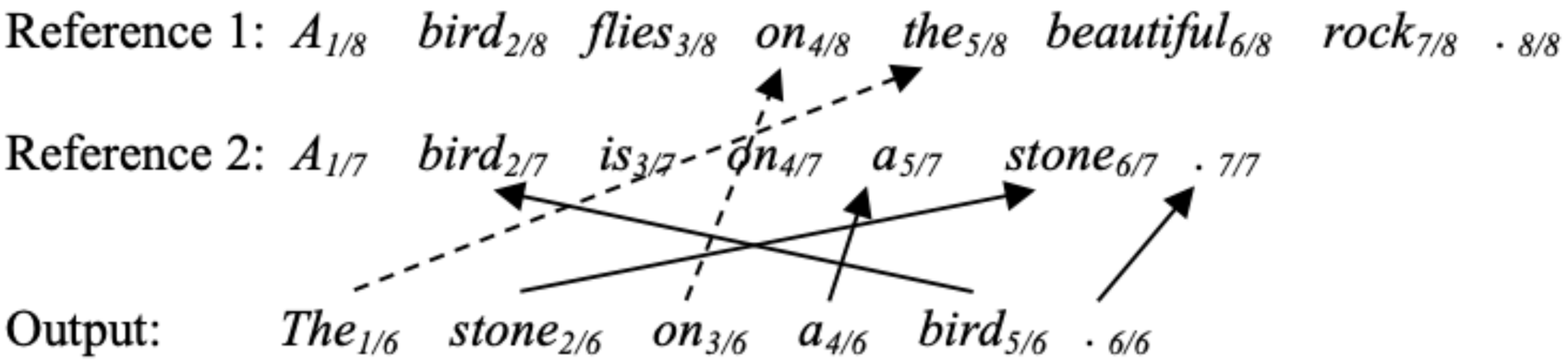
Single reference Example mapping *NPD* calculation





$$NPD = \frac{1}{6} \times \left[\left|\frac{1}{6} - \frac{5}{7}\right| + \left|\frac{2}{6} - \frac{6}{7}\right| + \left|\frac{3}{6} - \frac{4}{7}\right| + \left|\frac{4}{6} - \frac{1}{7}\right| + \left|\frac{5}{6} - \frac{2}{7}\right|\right] = \frac{2}{7}$$

Multi-reference Example mapping
NPD calculation



The beginning output words "the" and "stone" are aligned simply for the single matching. The output word "on" has nearby matching with the word "on" both in Reference-1 and Reference-2, due to the words "the" (second to previous) and "a" (first in the following) respectively. Then we should select its alignment to the word "on" in Reference-1, not Reference-2 for the further reason $\left|\frac{3}{6}-\frac{4}{8}\right| < \left|\frac{3}{6}-\frac{4}{7}\right|$ and this selection will obtain a smaller *NPD* value. The remaining two words "a" and "bird" in output sentence are aligned using the same principle.

Evaluation system	Correlation Score with Human Judgment								
	other-to-English				English-to-other				Mean score
	CZ-EN	DE-EN	ES-EN	FR-EN	EN-CZ	EN-DE	EN-ES	EN-FR	
LEPOR-B	0.93	0.62	0.96	0.89	0.71	0.36	0.88	0.84	**0.77**
LEPOR-A	0.95	0.61	0.96	0.88	0.68	0.35	0.89	0.83	**0.77**
AMBER	0.88	0.59	0.86	0.95	0.56	0.53	0.87	0.84	0.76
Meteor-1.3-RANK	0.91	0.71	0.88	0.93	0.65	0.30	0.74	0.85	0.75
BLEU	0.88	0.48	0.90	0.85	0.65	0.44	0.87	0.86	0.74
TER	0.83	0.33	0.89	0.77	0.50	0.12	0.81	0.84	0.64
MP4IBM1	0.91	0.56	0.12	0.08	0.76	0.91	0.71	0.61	0.58

Testing on WMT2011 data: correlation to HumanEval

Example of using POS for MTE

$$Harmonic(w_{X_1}X_1, w_{X_2}X_2, ..., w_{X_n}X_n) =$$

$$\frac{\sum_{i=1}^{n} w_{X_i}}{\sum_{i=1}^{n} \frac{w_{X_i}}{X_i}}$$

$$hLEPOR =$$

$$= \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} \frac{w_i}{Factor_i}} = \frac{w_{ELP} + w_{NPosPenal} + w_{HPR}}{\frac{w_{ELP}}{ELP} + \frac{w_{NPosPenal}}{NPosPenal} + \frac{w_{HPR}}{HPR}}$$

$$\overline{hLEPOR_E} = \frac{1}{w_{hw} + w_{hp}}(w_{hw}\overline{hLEPOR_{word}}$$

$$+w_{hp}\overline{hLEPOR_{POS}}) \qquad (11)$$

- LP: length penalty
- NPosPenal: n-gram based position difference penalty
- R: recall
- P: precision
- LEPOR(A): system-level A
- LEPOR(B): system-level B
- ELP: to distinguish from BLEU
- (Enhanced LP)
- hLEPOR: hybrid LEPOR

Han et al. (2013) "Language-independent Model for Machine Translation Evaluation with Reinforced Factors" in MT Summit. https://aclanthology.org/2013.mtsummit-posters.3.pdf Git https://github.com/poethan/aaron-project-hlepor

Metrics	Other-to-English				English-to-Other				Mean
	CZ-EN	DE-EN	ES-EN	FR-EN	EN-CZ	EN-DE	EN-ES	EN-FR	
$hLEPOR_E$	0.93	0.86	0.88	0.92	0.56	0.82	0.85	0.83	**0.83**
MPF	0.95	0.69	0.83	0.87	0.72	0.63	0.87	0.89	0.81
ROSE	0.88	0.59	0.92	0.86	0.65	0.41	0.9	0.86	0.76
METEOR	0.93	0.71	0.91	0.93	0.65	0.3	0.74	0.85	0.75
BLEU	0.88	0.48	0.9	0.85	0.65	0.44	0.87	0.86	0.74
TER	0.83	0.33	0.89	0.77	0.5	0.12	0.81	0.84	0.64

Testing on WMT2013 data: Correlation to Human Eval

hLEPOR @ WMT13

Directions	EN-FR	EN-DE	EN-ES	EN-CS	EN-RU	Av
LEPOR_v3.1	*.91*	*.94*	*.91*	*.76*	***.77***	***.86***
nLEPOR_baseline	*.92*	*.92*	*.90*	***.82***	*.68*	*.85*
SIMP-BLEU_RECALL	**.95**	.93	.90	**.82**	.63	.84
SIMP-BLEU_PREC	.94	.90	.89	**.82**	.65	.84
NIST-mteval-inter	.91	.83	.84	.79	.68	.81
Meteor	.91	.88	.88	**.82**	.55	.81
BLEU-mteval-inter	.89	.84	.88	.81	.61	.80
BLEU-moses	.90	.82	.88	.80	.62	.80
BLEU-mteval	.90	.82	.87	.80	.62	.80
CDER-moses	.91	.82	.88	.74	.63	.80
NIST-mteval	.91	.79	.83	.78	.68	.79
PER-moses	.88	.65	.88	.76	.62	.76
TER-moses	.91	.73	.78	.70	.61	.75
WER-moses	.92	.69	.77	.70	.61	.74
TerrorCat	.94	**.96**	**.95**	na	na	.95
SEMPOS	na	na	na	.72	na	.72
ACTa	.81	-.47	na	na	na	.17
ACTa5+6	.81	-.47	na	na	na	.17

Table 3. System-level Pearson correlation scores on WMT13 English-to-other language pairs

Han et al. (2013)
A Description of Tunable Machine Translation Evaluation Systems in WMT13 Metrics Task

In Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 414–421.
https://aclanthology.org/www.mt-archive.info/10/WMT-2013-Han-2.pdf

Example of using syntactic chunk

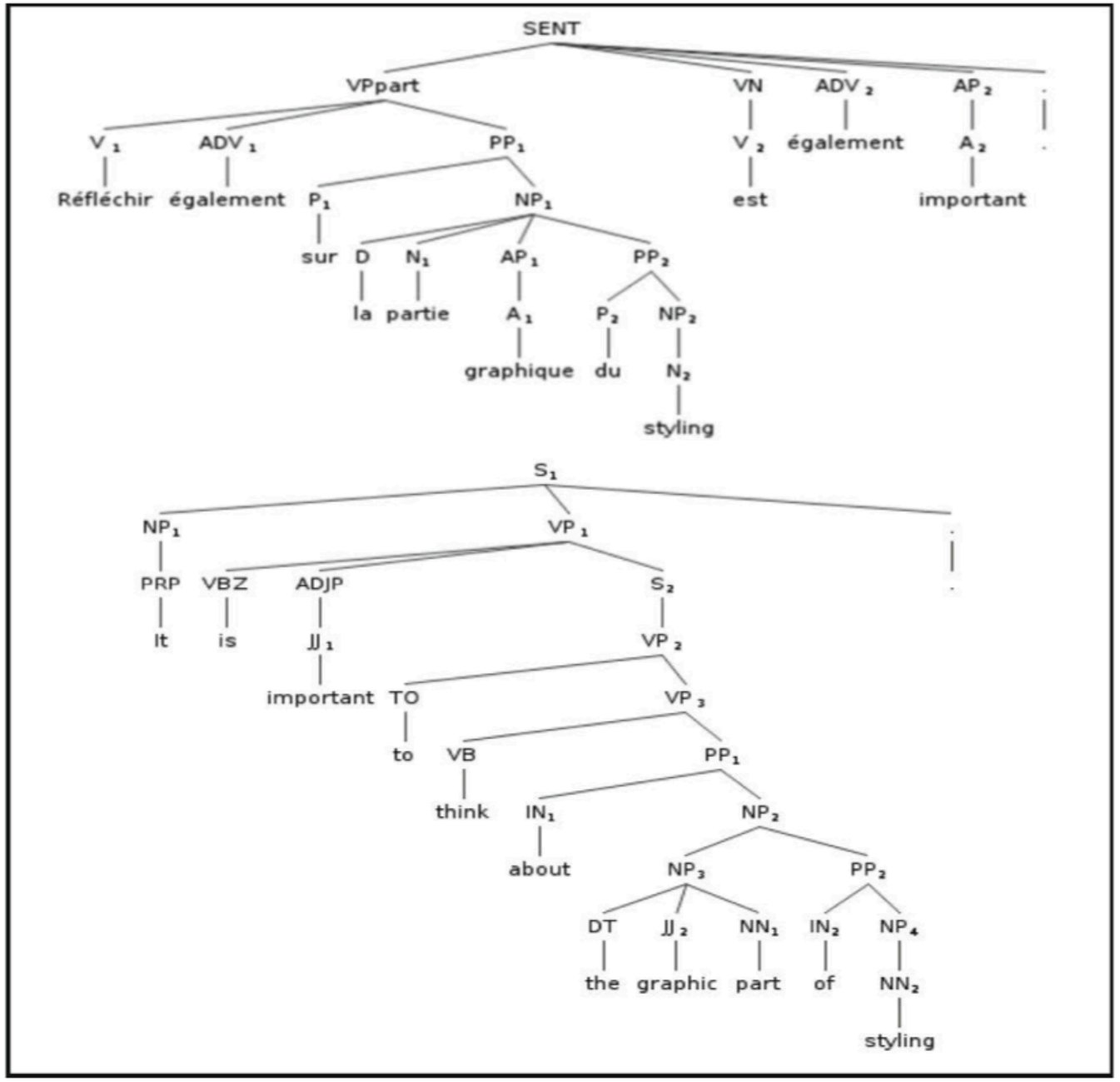
$$HPPR = Har(w_{Ps}\overline{N_1 PsDif}, w_{Pr}N_2 Pre, w_{Rc}N_3 Rec)$$

$$= \frac{w_{Ps} + w_{Pr} + w_{Rc}}{\frac{w_{Ps}}{\overline{N_1 PsDif}} + \frac{w_{Pr}}{N_2 Pre} + \frac{w_{Rc}}{N_3 Rec}}$$

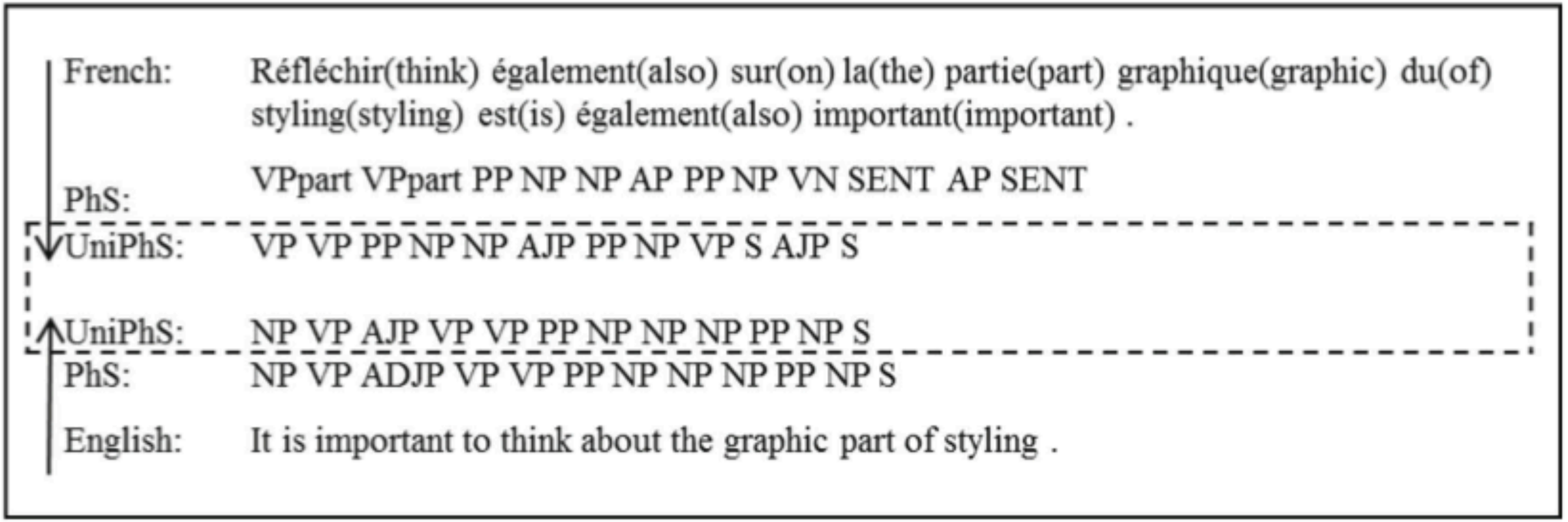
$$P_n = \frac{\#matched\ ngram\ chunks}{\#ngram\ chunks\ of\ hypothesis\ corpus}$$

$$R_n = \frac{\#matched\ ngram\ chunks}{\#ngram\ chunks\ of\ source\ corpus}$$

https://github.com/poethan/aaron-project-hppr



Han et al. (2013) "Phrase Tagset Mapping for French and English Treebanks and Its Application in Machine Translation Evaluation". GSCL2013. https://link.springer.com/content/pdf/10.1007/978-3-642-40722-2_13.pdf



Han et al. (2014) "A Universal Phrase Tagset for Multilingual Treebanks". NLP-NABD, CCL2014. https://link.springer.com/chapter/10.1007/978-3-319-12277-9_22

DL model distillation example for MTE

Our work cushLEPOR to customise hLEPOR towards LaBSE/HumanEval

- Pre-trained Language Models (PLMs) are costly regarding computation
- Human Eval is costly regarding time and money
- How to customise our existing metrics towards human performance
- How to costomise existing metrics towards PLM performance but with low cost
- We use our own metric hLEPOR, and customise it using hyper-parameter optimisation (Optuna framework) towards HumanEval / PLMs

Lifeng Han, Irina Sorokina, Gleb Erofeev, Serge Gladkoff (2021) cushLEPOR: customising hLEPOR metric using Optuna for higher agreement with human judgments or pre-trained language model LaBSE. WMT21. https://aclanthology.org/2021.wmt-1.109/
Gleb Erofeev, Irina Sorokina, Lifeng Han, Serge Gladkoff (2021) cushLEPOR uses LABSE distilled knowledge to improve correlation with human translation evaluations https://aclanthology.org/2021.mtsummit-up.28/

Our work cushLEPOR to customise hLEPOR towards LaBSE/HumanEval

- **alpha**: the tunable weight for recall
- **beta**: the tunable weight for precision
- **n**: words count before and after matched word in npd calculation
- **weight_elp**: tunable weight of enhanced length penalty
- **weight_pos**: tunable weight of n-gram position difference penalty
- **weight_pr**: tunable weight of harmonic mean of precision and recall

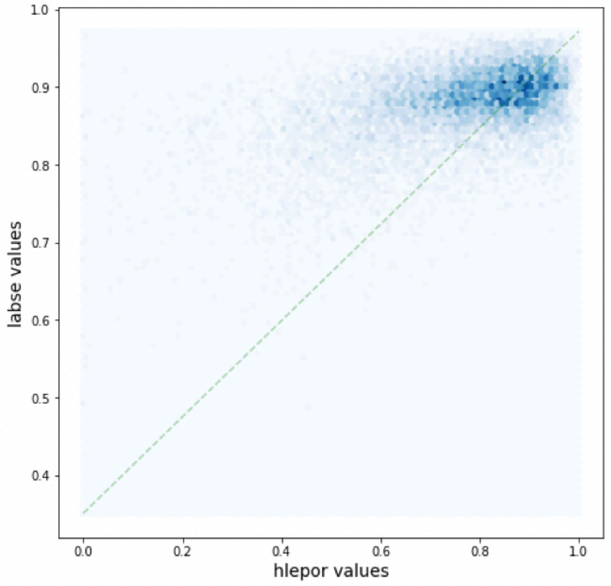


Figure 1: Agreement with LaBSE: hLEPOR

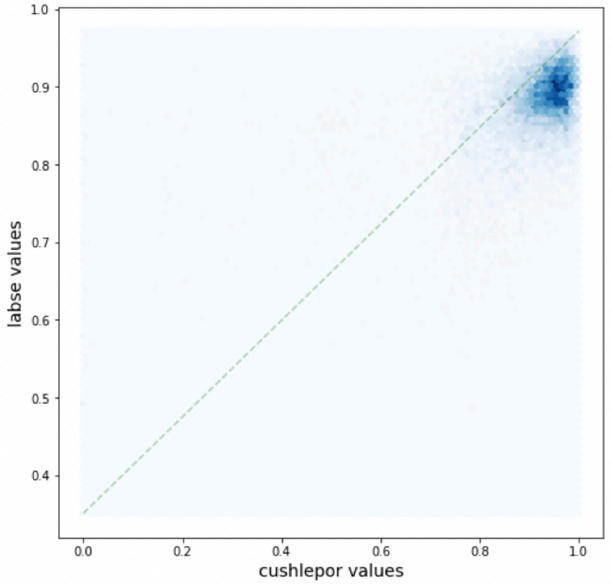


Figure 2: Agreement with LaBSE: cushLEPOR

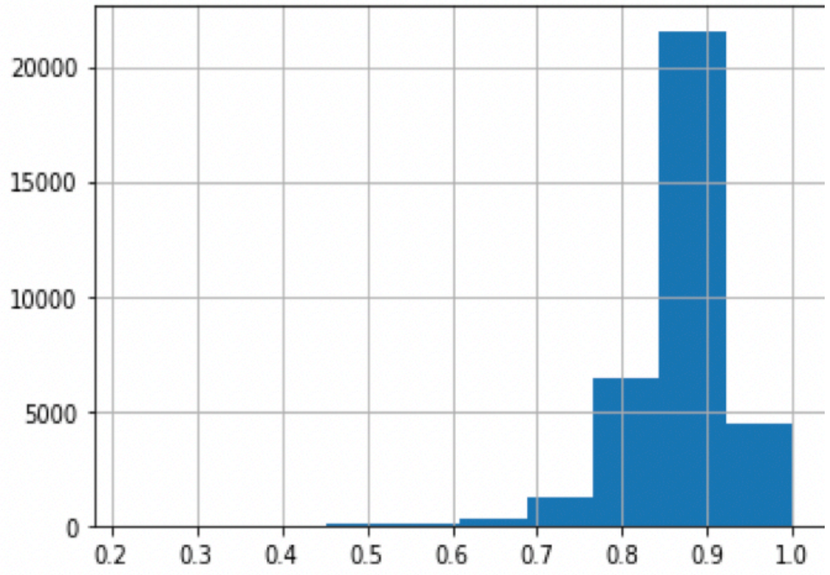


Figure 6: Score Distribution: tune on LaBSE

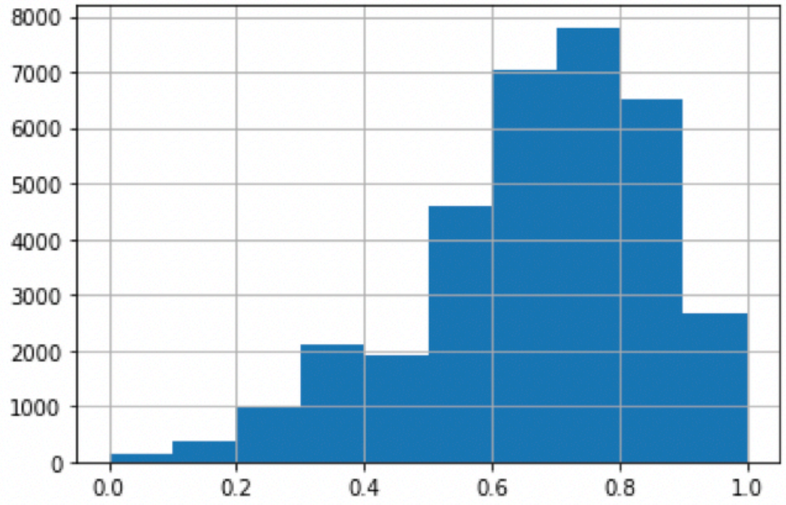


Figure 7: Score Distribution: tune on pSQM

Our work cushLEPOR to customise hLEPOR towards LaBSE / HumanEval (pSQM)

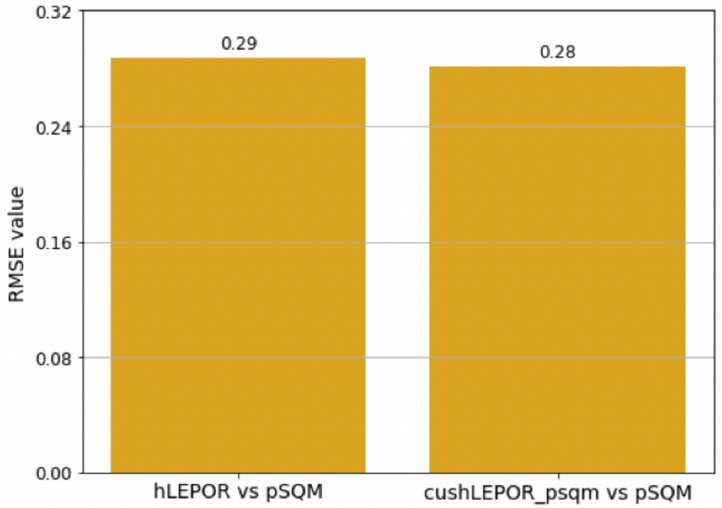


Figure 3: RMSE: hLEPOR vs cushLEPOR to pSQM (lower score is better)

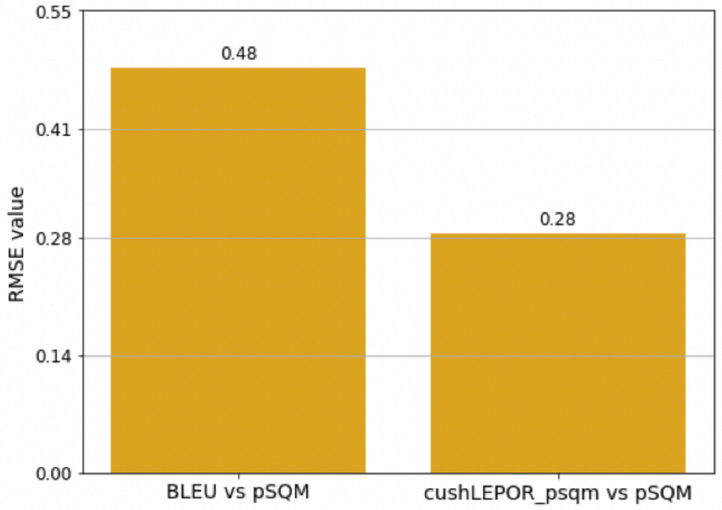


Figure 4: RMSE: BLEU vs cushLEPOR to pSQM (lower score is better)

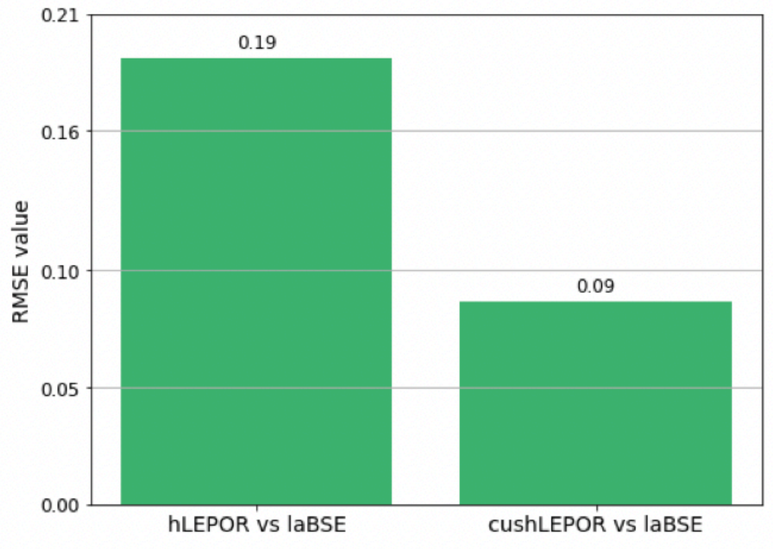


Figure 5: RMSE: hLEPOR vs cushLEPOR to LaBSE (lower score is better)

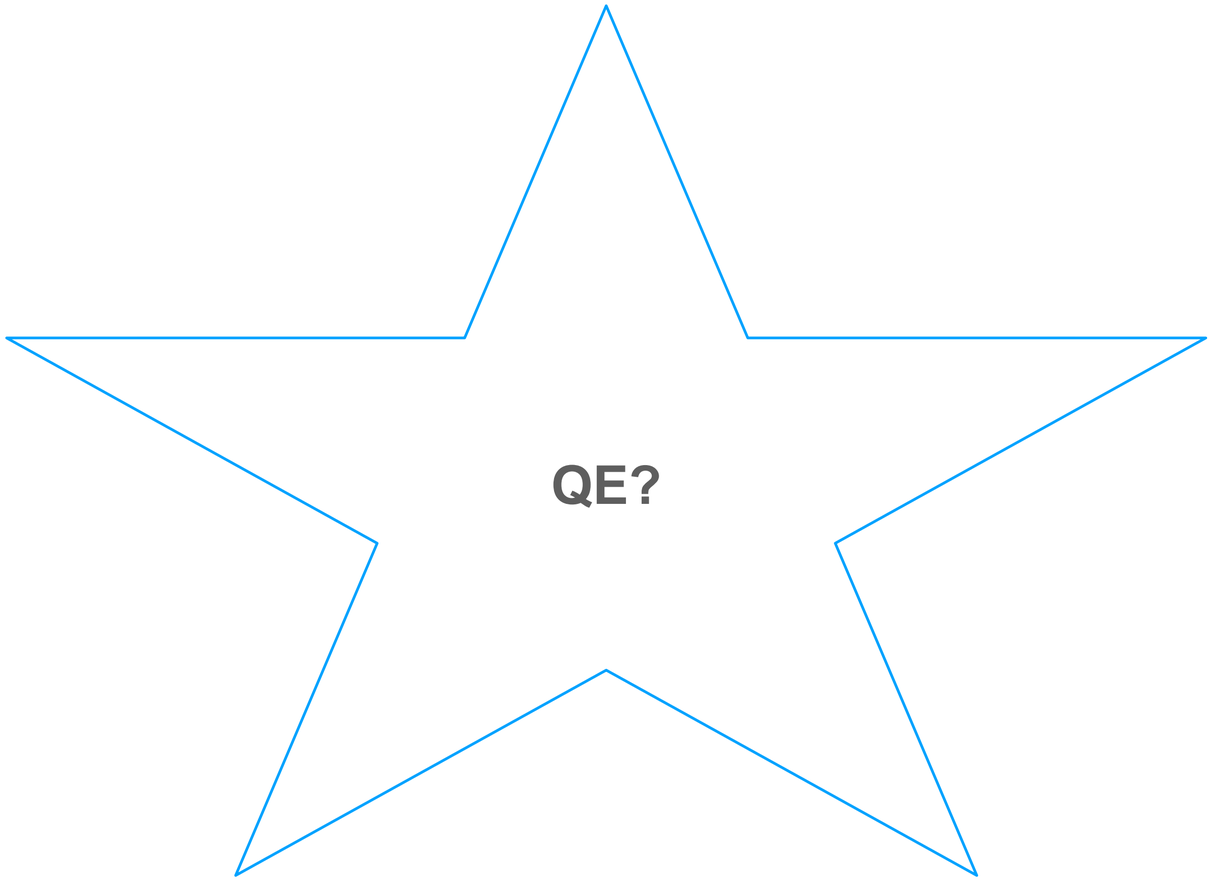
Our work cushLEPOR has much better performance in comparison to BLEU, regarding RMSE error (almost half reduction) towards professional HumanEval

The official results from WMT2021 Metrics task show that cushLEPOR(LM) ranks in the **first cluster** in performance on **News test** data with single reference evaluated on overall English-to-German, Chinese-to-English and English-to-Russian where professional human evaluation data is available (*Ref.* Table 8 "Metric rankings based on pairwise accuracy" in Findings paper (Freitag et al., 2021)). Furthermore, in the language specific ranking, cushLEPOR(LM) also wins **English-to-German** and **Chinese-to-English** language pairs, including TED data condition. Our hLEPOR baseline metric wins **English-to-Russian** TED domain language specific ranking (Ref. Table 12 "Summary of language-specific results" in the official findings paper (Freitag et al., 2021)). The official result on "System-level Pearson correlations for **English-to-German**" (Table 23 of findings) shows that cushLEPOR(LM) achieves score 0.938 in News domain, ranking *number 1 in Cluster 1 metrics*, out of overall 29 metric submissions.

Lifeng Han, Irina Sorokina, Gleb Erofeev, Serge Gladkoff (2021)

cushLEPOR: customising hLEPOR metric using Optuna for higher agreement with human judgments or pre-trained language model LaBSE. WMT21.
https://aclanthology.org/2021.wmt-1.109/

Python Code: https://pypi.org/project/hLepor/
Data/scoring: https://github.com/poethan/LEPOR


QE?

Automatic TQA: QE

How to estimate the translation quality without depending on the prepared reference translation?

- e.g. based on the features that can be extracted from MT output and source text
- estimation spans: word, sentence, system level.

Evaluation methods:

- word level: Mean Absolute Error (MAE) as a primary metric, and Root of Mean Squared Error (RMSE) as a secondary metric, borrowed from measuring regression task.
- sentence level: DA, HTER
- document level: MQM (multi-dimention quality metrics)

Automatic TQA: QE

Recent trends:

- Metrics and QE integration/cross performance, how to utilise/ communicate both metrics task and QE task knowledge

-WMT19: there were 10 reference-less evaluation metrics which were used for the QE task, “QE as a Metric”

- challenges: predicting src-word that leads error, multilingual or language independent models, low resource languages, etc.

- Automatic TQA: QE - our work

Statistical modelling using metrics criteria for QE

I.- sentence-level quality estimation: EBLEU (an enhanced BLEU with POS)

II.- system selection: probability model Naïve Bayes (NB) and SVM as classification algorithm with the features from evaluation metrics (ength penalty, Precision, Recall and Rank values)

III.- word-level quality estimation: take the contextual information into account, we employed a discriminative undirected probabilistic graphical model conditional random field (CRF), in addition to the NB algorithm.

Han et al. (2013) Quality Estimation for Machine Translation Using the Joint Method of Evaluation Criteria and Statistical Modeling. In WMT13. https://www.aclweb.org/anthology/W13-2245/

Meta eval - Evaluating MTE credibility

Statistical **significance**: by bootstrap re-sampling methods, different system indeed own different quality? statistical significance intervals for evaluation metrics on small test sets.

Human judgement **agreement** *kappa* levels: how often they agree with each other on the segments, or they agree with themselves on the same segment.

Correlation methods: how much correlation is the candidate ranking and scoring (metrics, QE), compared to the reference ranking (e.g. human)? *Spearman, Pearson, Kendall tau*.

Metrics comparisons: compare and manipulate different metrics.

HumanEval and **Test set** validation, confidence

Human eval agreement

Since the human judgments are usually trusted as the golden standards that the automatic evaluation metrics should try to approach, the reliability and coherence of human judgments is very important. Cohen's kappa agreement coefficient is one of the commonly used evaluation methods (Cohen, 1960). For the problem in nominal scale agreement between two judges, there are two relevant quantities p_0 and p_c. The factor p_0 is the proportion of units in which the judges agreed and p_c is the proportion of units for which agreement is expected by chance. The coefficient k is simply the proportion of chance-expected disagreements which do not occur, or alternatively, it is the proportion of agreement after chance agreement is removed from consideration:

$$k = \frac{p_0 - p_c}{1 - p_c} \tag{14}$$

where $p_0 - p_c$ represents the proportion of the cases in which beyond-chance agreement occurs and is the numerator of the coefficient (Landis and Koch, 1977).

Correlation coefficients

4.3.1 *Pearson Correlation*

Pearson's correlation coefficient (Pearson, 1900) is commonly represented by the Greek letter ρ. The correlation between random variables X and Y denoted as is measured as follow (Montgomery and Runger, 2003).

$$\rho_{XY} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{15}$$

Because the standard deviations of variable X and Y are higher than 0 ($\sigma_X > 0$ and $\sigma_Y > 0$), if the covariance σ_{XY} between X and Y is positive, negative or zero, the correlation score between X and Y will correspondingly result in positive, negative or zero, respectively. Based on a sample of paired data (X, Y) as $(x_i, y_i), i = 1 \, to \, n$, the Pearson correlation coefficient is calculated by:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{n}(x_i - \mu_x)^2}\sqrt{\sum_{i=1}^{n}(y_i - \mu_y)^2}} \tag{16}$$

where μ_x and μ_y specify the means of discrete random variable X and Y respectively.

Correlation coefficients

4.3.2 *Spearman rank Correlation*

Spearman rank correlation coefficient, a simplified version of Pearson correlation coefficient , is another algorithm to measure the correlations of automatic evaluation and manual judges, especially in recent years (Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010; Callison-Burch et al., 2011). When there are no ties, Spearman rank correlation coefficient, which is sometimes specified as (rs) is calculated as:

$$rs_{\varphi(XY)} = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} \qquad (17)$$

where d_i is the difference-value (D-value) between the two corresponding rank variables $(x_i - y_i)$ in $\vec{X} = \{x_1, x_2, ..., x_n\}$ and $\vec{Y} = \{y_1, y_2, ..., y_n\}$ describing the system φ.

Correlation coefficients

$$\tau = \frac{\text{num concordant pairs} - \text{num discordant pairs}}{\text{total pairs}} \tag{18}$$

The latest version of Kendall's τ is introduced in (Kendall and Gibbons, 1990). (Lebanon and Lafferty, 2002) give an overview work for Kendall's τ showing its application in calculating how much the system orders differ from the reference order. More concretely, (Lapata, 2003) proposes the use of Kendall's τ, a measure of rank correlation, estimating the distance between a system-generated and a human-generated gold-standard order.

Meta-eval - example

Google research on WMT (2020) data credibility

Computer Science > Computation and Language

[Submitted on 29 Apr 2021]

Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, Wolfgang Macherey

WMT conducted a reference-based/monolingual human evaluation for Chinese→English in which the machine translation output was compared to a human-generated reference. When comparing the system ranks based on WMT for both language pairs with the ones generated by MQM, we can see low correlation for English→German (see Figure 1) and even negative correlation for Chinese→English (see Figure 3). We also see very

- Freitag et al. (2021) https://arxiv.org/abs/2104.14478
- Meta-eval on WMT2020 data

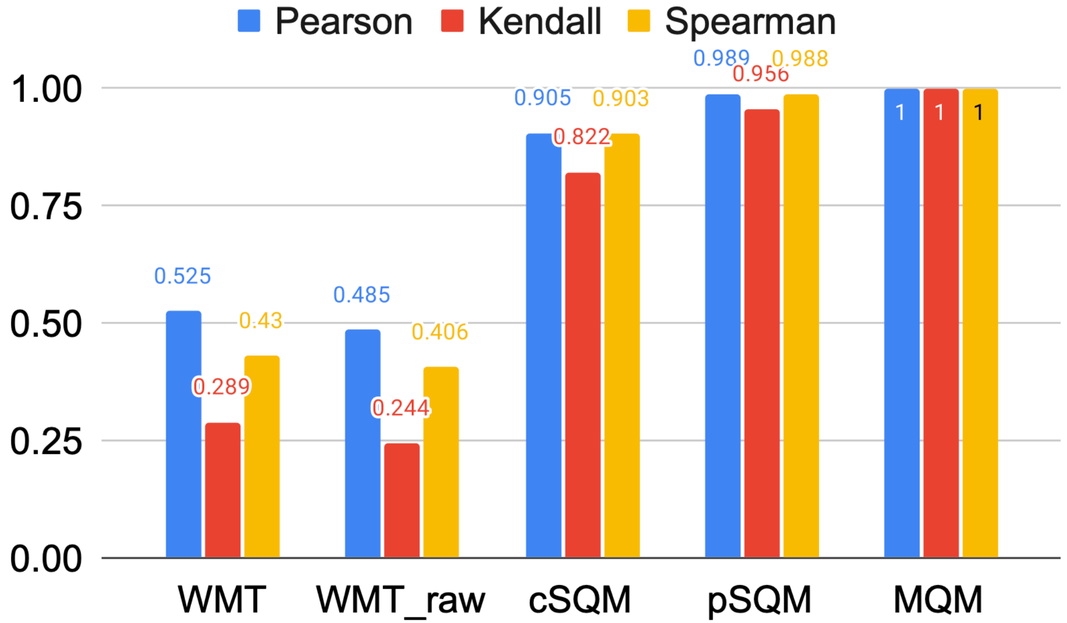


Figure 1: English→German: System correlation with the platinum ratings acquired with MQM.

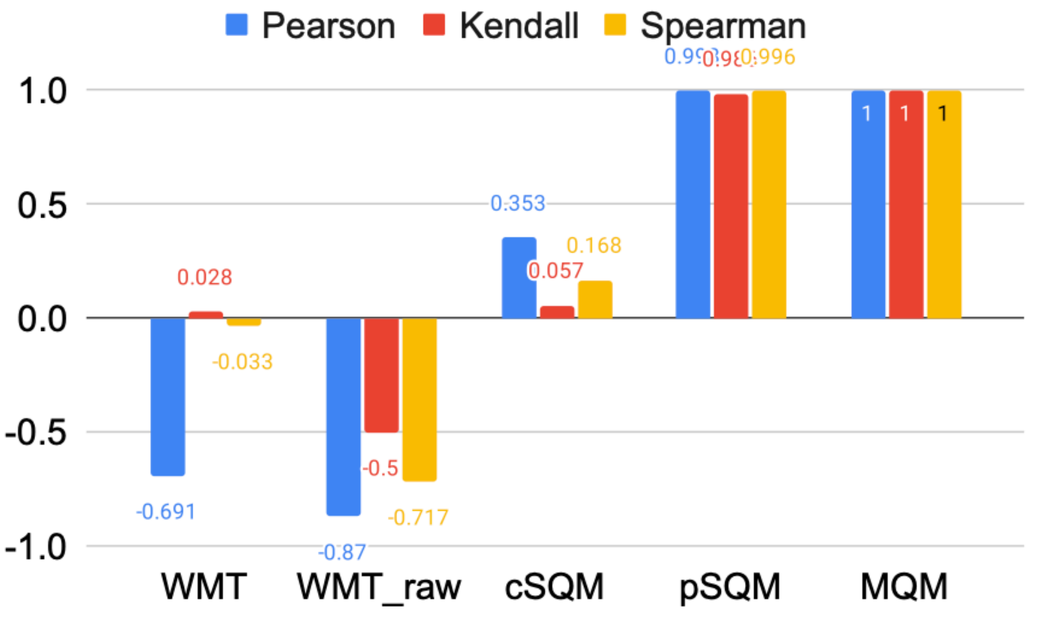


Figure 3: Chinese→English: System-level correlation with the platinum ratings acquired with MQM.

(i) Human translations are underestimated by crowd workers: Already in 2016, Hassan et al. (2018) claimed human parity for news-translation for Chinese→English. We confirm the findings of Toral et al. (2018); Läubli et al. (2018) that when human evaluation is conducted correctly, professional translators can discriminate between human and machine translations. All human translations

(i) Human translations are underestimated by crowd workers: Already in 2016, Hassan et al. (2018) claimed human parity for news-translation for Chinese→English. We confirm the findings of Toral et al. (2018); Läubli et al. (2018) that when human evaluation is conducted correctly, professional translators can discriminate between human and machine translations. All human translations

- Freitag et al. (2021) https://arxiv.org/abs/2104.14478
- Meta-eval on WMT2020 data

Meta-eval - example

unveil the model flaws from published work

- Check-List by Ribeiro et al. (2022) " Beyond Accuracy: Behavioral Testing of NLP Models with CheckList" -> comprehensive behavioral testing of NLP models
- - a list of linguistic ‘capabilities’: vocabulary, NER, Negation, semantic role labelling (agent, object), sentiment, POS, Temporal
- - Logic: ability to handle symmetry, consistency, and conjunctions.
- - break down potential capability failures into specific behaviours: robustness to typos, irrelevant changes
- - different ‘test types’, e.g. sanity checks

- Ribeiro et al. (2022) "ACL-2020: Beyond Accuracy: Behavioral Testing of NLP Models with CheckList"

Meta-eval - example

Their findings Ribeiro et al. (2022)

- Sentiment on DL based models: BERT-base, RoBERTa-base
- - using SST-2 (Stanford Sentiment Treebank) data
- - BERT and RoBERTa do poorly on neutral predictions (being trained on binary)
- - none do well on Temporal, Negation, SRL capabilities
- - unveil data-set bias: BERT-base “always predicts negative when {protected} is black, atheist, gay and lesbian, while predicting positive for Asian, straight”

Meta-eval - example

Their findings Ribeiro et al. (2022)

- On QQP data: quota question pair
- - both DL models “ignoring important modifiers on the *Vocab.* test, and lacking basic *Taxonomy* under- standing, e.g. synonyms and antonyms of common words.”
- - *SRL* tests, neither model is able to handle agent/predicate changes
- - models often fail to make simple *Temporal* distinctions (e.g. is *vs* used to be and before *vs* after), and to distinguish between simple *Coreferences* (he *vs* she).

.

Human Parity?

Hassan et al. (2020)
- Really?! I do not need to finish my PhD anymore

Achieving Human Parity on Automatic
Chinese to English News Translation

Hany Hassan,* Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark,
Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis,
Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin,
Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia,
Dongdong Zhang, Zhirui Zhang, and Ming Zhou

Microsoft AI & Research

- https://arxiv.org/abs/1803.05567

Hassan et al. (2020)

Intuitively, we can define human parity for translation as follows:

Definition 1. *If a bilingual human judges the quality of a candidate translation produced by a human to be equivalent to one produced by a machine, then the machine has achieved* HUMAN PARITY.

Assuming that it is possible for humans to measure translation quality by assigning scores to translations of individual sentences of a test set, and generalizing from a single sentence to a set of test sentences, this effectively yields the following statistical definition:

Definition 2. *If there is no statistically significant difference between human quality scores for a test set of candidate translations from a machine translation system and the scores for the corresponding human translations then the machine has achieved* HUMAN PARITY.

the current NMT paradigm. Our contributions are:

- We utilize the duality of the translation problem to allow the model to learn from both source-to-target and target-to-source translations. Simultaneously this allows us to learn from both supervised and unsupervised source and target data. This will be described in Section 3.3. Specifically, we utilize a generic Dual Learning approach [20, 48, 47] (Section 3.3.1), and introduce a joint training algorithm to enhance the effect of monolingual source and target data by iteratively boosting the source-to-target and target-to-source translation models in a unified framework (Section 3.3.2).

- NMT systems decode auto-regressively from left-to-right, which means that during sequential generation of the output, previous errors will be amplified and may mislead subsequent generation. This is only partially remedied by beam search. We propose two approaches to alleviate this problem: Deliberation Networks [49] is a method to refine the translation based on two-pass decoding (Section 3.4.1); and a new training objective over two Kullback-Leibler (KL) divergence regularization terms encourages agreement between left-to-right and right-to-left decoding results (Section 3.4.2).

- Since NMT is very vulnerable to noisy training data, rare occurrences in the data, and the training data quality in general [4]. We discuss our approaches for data selection and filtering, including a cross-lingual sentence representation, in Section 3.5.

- Finally, we find that our systems are quite complementary, and can therefore benefit greatly from system combination, ultimately attaining human parity. See section 3.6.

In this work, we interchangeably use source-to-target and (Zh→En) to denote Chinese-to-English; target-to-source and (En→Zh) to denote English-to-Chinese.

Hassan et al. (2020)

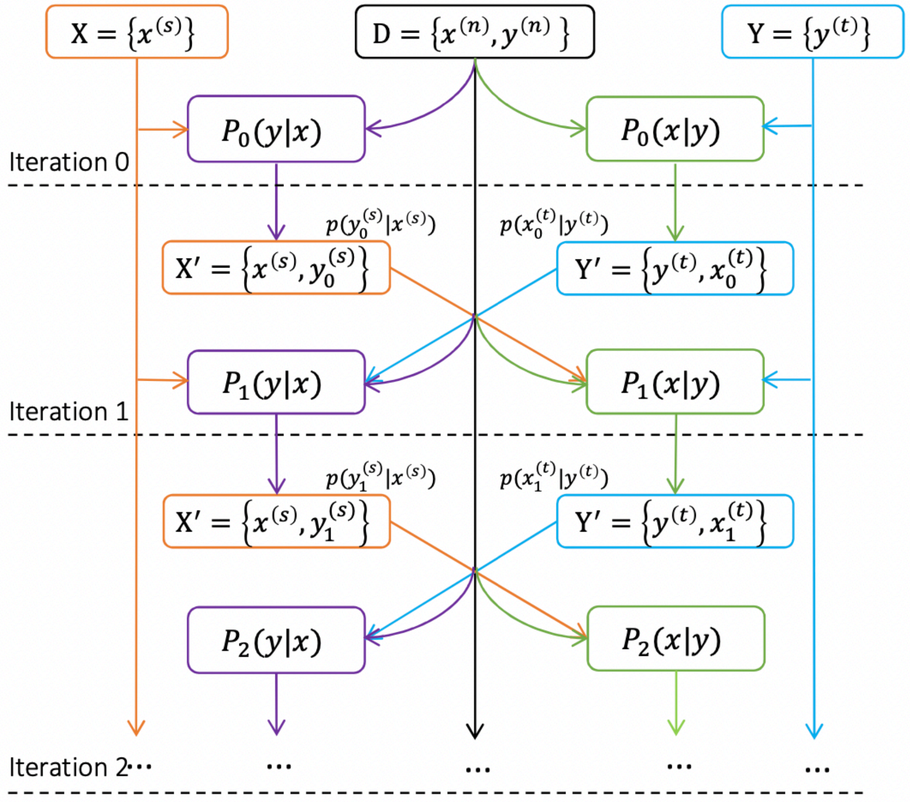


Figure 1: Illustration of joint training: S2T $p(\mathbf{y}|\mathbf{x})$ and T2S $p(\mathbf{x}|\mathbf{y})$

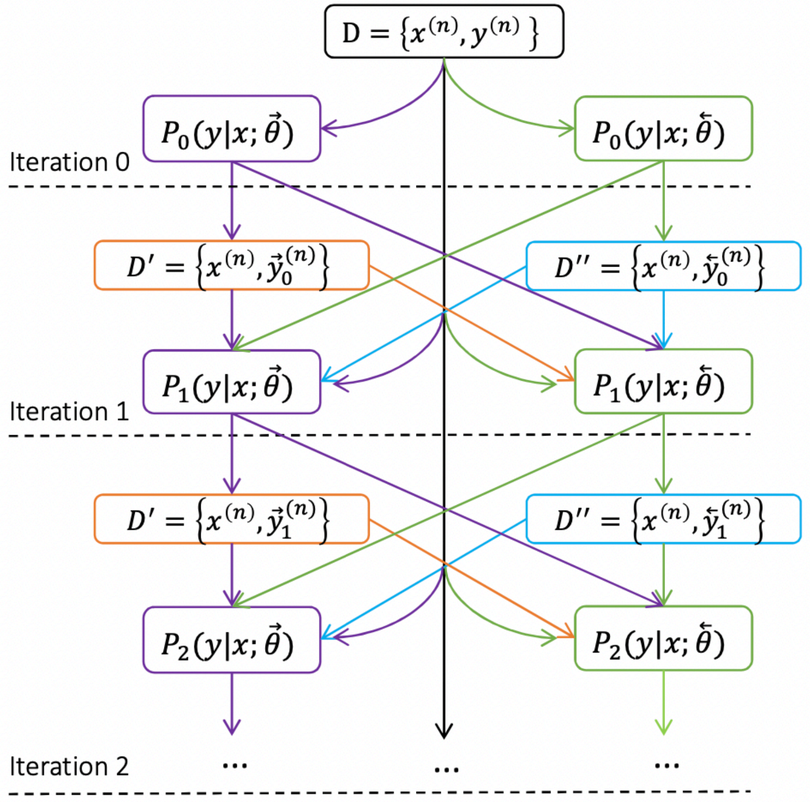
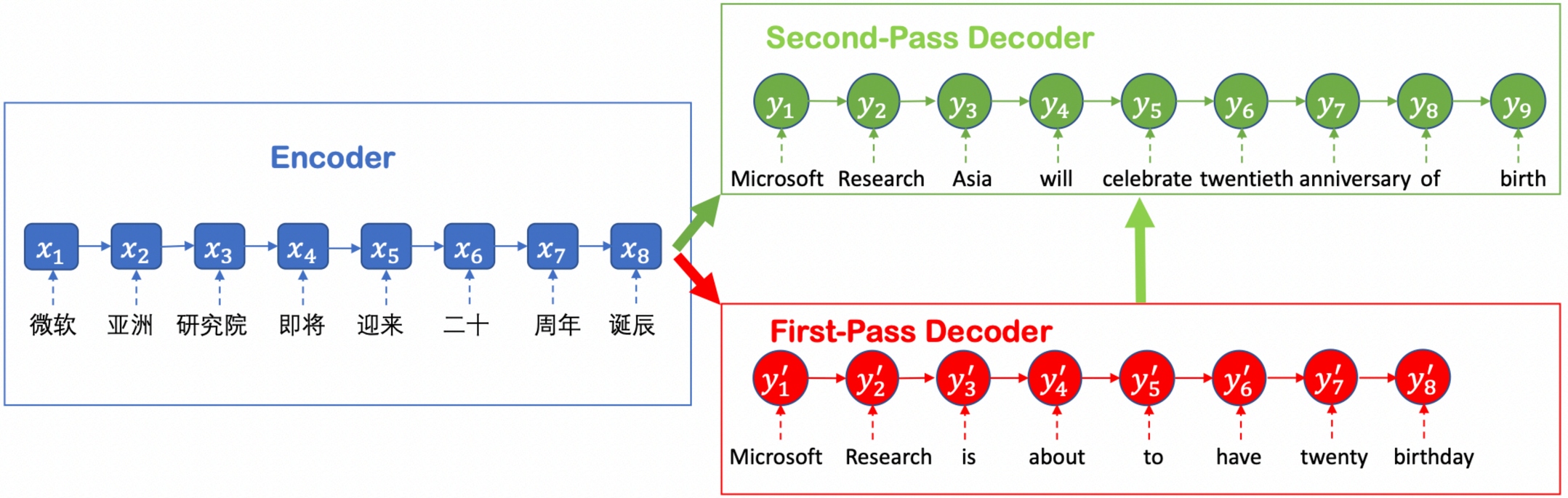


Figure 2: Illustration of agreement regularization: L2R $p(\mathbf{y}|\mathbf{x};\overrightarrow{\theta})$ and R2L $p(\mathbf{y}|\mathbf{x};\overleftarrow{\theta})$

Hassan et al. (2020)



Hassan et al. (2020)

Meta-eval - “the human parity”

Läubli et al.(2020) findings

- “reassess Hassan et al.’s 2018 investigation into Chinese to English news translation, showing that the finding of human–machine parity was owed to weaknesses in the **evaluation design**”

- “Achieving Human Parity on Automatic Chinese to English News Translation" https://github.com/MicrosoftTranslator/Translator-HumanParityData

- “We show that the professional human translations contained significantly fewer errors, and that perceived quality in human evaluation depends on the choice of **raters**, the availability of **linguistic context**, and the creation of **reference** translations. Our results call for revisiting current best practices to assess strong machine translation systems in general and human–machine parity in particular, for which we offer a set of recommendations based on our empirical findings.”

Läubli et al.(2020) A set of recommendations for assessing human-machine parity in language translation https://jair.org/index.php/jair/article/view/11371

Errors in Understanding: If the translator is a non-native speaker of the source language, they may make mistakes in interpreting the original message. This is particularly true if the translator does not normally work in the domain of the text, e. g., when a translator who normally works on translating electronic product manuals is asked to translate news.

Errors in Fluency: If the translator is a non-native speaker of the target language, they might not be able to generate completely fluent text. This similarly applies to domain-specific terminology.

Limited Resources: Unlike computers, human translators have limits in time, attention, and motivation, and will generally do a better job when they have sufficient time to check their work, or are particularly motivated to do a good job, such as when doing a good job is necessary to maintain their reputation as a translator.

Effects of Post-editing: In recent years, a large number of human translation jobs are performed by post-editing MT output, which can result in MT artefacts remaining even after manual post-editing (Castilho, Resende, & Mitkov, 2019; Daems, Vandepitte, Hartsuiker, & Macken, 2017; Toral, 2019).

Läubli et al.(2020) findings

Source	传统习俗引入新亮点“**2016孟兰文化节**”香港维园开幕敲锣打鼓的音乐、传统的小食、花俏的装饰、人群汹涌的现场。由香港潮属社团总会主办的“**2016孟兰文化节**”12日至14日在维多利亚公园举办，这是香港最盛大的一场盂兰胜会。
H_A	Traditional customs with new highlights - **2016 Ullam Cultural Festival** unveiled at Victoria Park in Hong Kong Music with drums and gongs, traditional snacks, fanciful decorations, and a chock-a-block crowd at the scene. The “**2016 Ullam Cultural Festival**” organized by the Federation of Hong Kong Chiu Chow Community Organizations will be held at Victoria Park from 12th to the 14th.
MT_1	Traditional customs introduce new bright spot “**2016 Ullambana Cultural Festival**” Hong Kong Victoria Park opening Gongs and drums music, traditional snacks, fancy decorations, the crowd surging scene. Organised by the Federation of Teochew Societies in Hong Kong, the “**2016 Python Cultural Festival**” is held at Victoria Park from 12 to 14 July.

MT1: Hassan et al. (2018)

Läubli et al.(2020) findings

(R1) Choose professional translators as raters. In our blind experiment (Section 3), non-experts assess parity between human and machine translation where professional translators do not, indicating that the former neglect more subtle differences between different translation outputs.

(R2) Evaluate documents, not sentences. When evaluating sentences in random order, professional translators judge machine translation more favourably as they cannot identify errors related to textual coherence and cohesion, such as different translations of the same product name. Our experiments show that using whole documents (i. e., full news articles) as unit of evaluation increases the rating gap between human and machine translation (Section 4).

(R3) Evaluate fluency in addition to adequacy. Raters who judge target language fluency without access to the source texts show a stronger preference for human translation than raters with access to the source texts (Sections 4 and 5.1). In all of our experiments, raters prefer human translation in terms of fluency while, just as in Hassan et al.'s (2018) evaluation, they find no significant difference between human and machine translation in sentence-level adequacy (Tables 2 and 4a). Our error analysis in Table 6 also indicates that MT still lags behind human translation in fluency, specifically in grammaticality.

(R4) Do not heavily edit reference translations for fluency. In professional translation workflows, texts are typically revised with a focus on target language fluency after an initial translation step. As shown in our experiment in Section 5.1, aggressive revision can make translations more fluent but less accurate, to the degree that they become indistinguishable from MT in terms of accuracy (Table 4a).

(R5) Use original source texts. Raters show a significant preference for human over machine translations of texts that were originally written in the source language, but not for source texts that are translations themselves (Section 5.2). Our results are further evidence that translated texts tend to be simpler than original texts, and in turn easier to translate with MT.

Läubli et al.(2020) findings

Meta-eval: our TQE uncertainty measuring

Gladkoff et al. (2022) LREC

- Confidence evaluation
- How large is large enough regarding testing set?
- How confident is the evaluation?
- How to simulate the evaluation quality from computational aspect?
- How to avoid bias, random judgment, and uncertainty?
- - Bernoulli statistical modeling and Monte Carlo Sampling Analysis

S Gladkoff, I Sorokina, L Han, A Alekseeva (2022) Measuring Uncertainty in Translation Quality Evaluation (TQE). LREC22. arXiv preprint arXiv:2111.07699

3.1 Study Setup

To carry out a statistical modeling of our research questions under study, i.e. *how to confidently choose a sample size of translation outputs to estimate the overall translation quality, from either HT or MT*, we setup the following initial assumptions:

- Translation errors belong to several independent categories
- Errors of one category are independent from each other.
- Errors of one category occur in text randomly.
- The smallest unit of text where language error occurs is a sentence (in other words, error can be between words, but cannot be between sentences).
- The resulting error distribution is a superposition of distribution of errors of different categories.

Then, if we further project these assumptions into simpler mathematical notes as bellow, it meets definition of Bernoulli trial (Brown et al., 2001; Agresti and Coull, 1998):

- Errors of certain type (category) either present in a sentence, or not.
- Errors are independent from each other.
- The probability of errors is the same.

3.2 Bernoulli Distribution

In Bernoulli statistical distribution, when the sample size n is significantly smaller than the overall population N, the standard derivation of sample measurement falls into the following formula:

$$\sigma = \sqrt{\frac{p \cdot (1-p)}{n}}$$

where p is the probability estimation of an event under study. The confidence interval CI, using the Wald interval (Newcombe, 2012)), will be:

$$CI = p \pm \Delta$$

where Δ is the product of standard deviation and factor 1.96 (when confidence level 95% is chosen) (Agresti and Coull, 1998).

$$\Delta = 1.96 \cdot \sigma = 1.96 \cdot \sqrt{\frac{p \cdot (1-p)}{n}}$$

When the sample size n is comparable to the population size N, e.g. in a smaller translation evaluation project for our study, the standard deviation is calculated as bellow and the Δ value updates correspondingly:

$$\sigma = \sqrt{\frac{p \cdot (1-p)}{n} \cdot \frac{N-n}{N-1}}$$

Let's come back to our study, with this even distribution assumption of each sentence regarding translation errors, having error probability p with value 1 and no error probability $1-p$ with value 0, each sentence represents a random variable in the modelling. This forms a Binomial distribution *B(n, p)* where *n* is the number of sentences.

3.3 Case Studies

We present case studies using both high quality translation text and low quality one. We first carried out a case study using high quality translation. Statistics from language service providers [ii] shows that the average length of English sentence is 17 words; there are about 250 words on standard page; we therefore can assume that there are 15 English sentences on a standard page. Let's assume that there is very high quality translation document, where there are no more than one error per page (one error per 15 sentences); then error density $p = 0.07$. And, if we set Δ value as 0.02. Thus the confidence interval falls into 0.07 ± 0.02, i.e. from 0.05 to 0.09, which is already wide interval. If we use a confidence level 95%, we have the following recommended number of sentences to check as derived from the formula mentioned earlier:

$$n = \frac{1.96^2 \cdot p \cdot (1-p)}{\Delta^2}$$

$$n = \frac{1.96^2 \cdot 0.07 \cdot (1-0.07)}{0.02^2} = 625$$

[ii] for instance, https://logrusglobal.com/

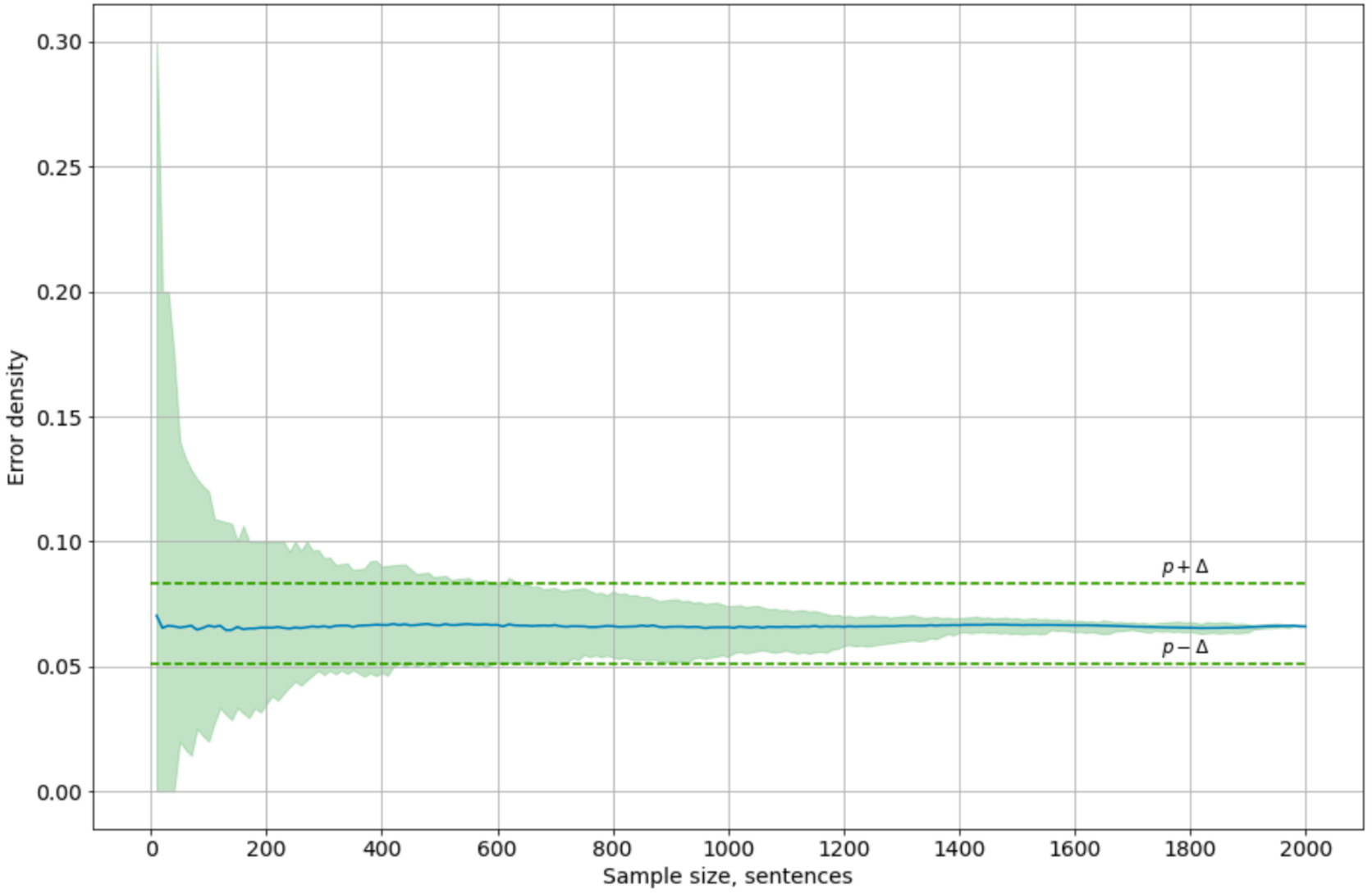


Fig. 1: A 95% confidence level credible interval for sample size from 100 to 2000 sentences, for high quality translation with error density 0.07. Δ is shown for sample size 625 sentences (42 pages).

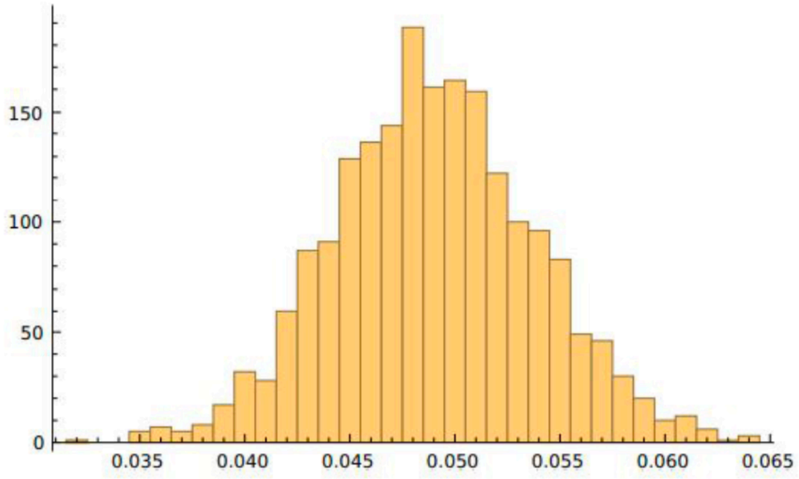


Fig. 3: Distribution histogram of error number in samples, with sample size of 1000 sentences.

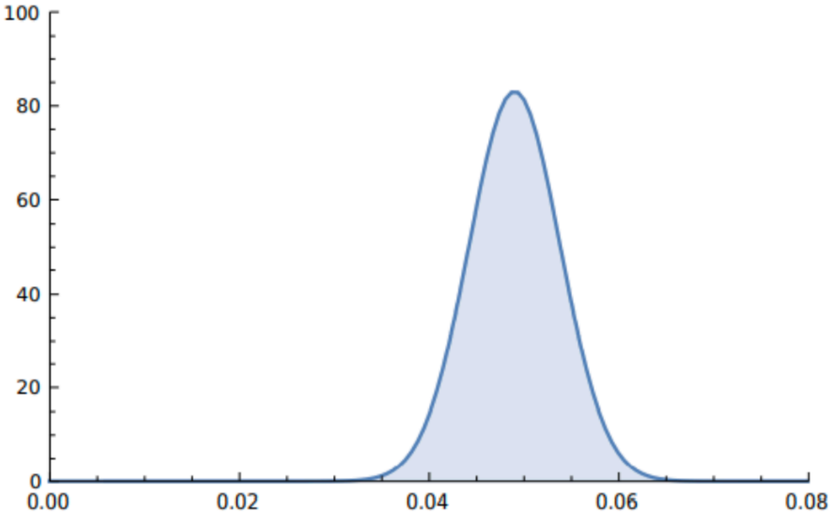


Fig. 4: Reconstructed probability density function of distribution of error number in samples of 1000 sentences.

Monte Carlo Simulation

The estimation of rare event probabilities are best analyzed with MCS method, sometimes it is probably the only way to handle such phenomena. As we mentioned, MCS relies on repeated random sampling to obtain numerical results for random variables where models are not available analytically. In our study, it is the case of translation error in the text, with many types of it which interact in a complex way, and the distribution of these errors is not uniform due to the text having a structure of unequal content. The possible reasons for this include that different people have worked on the entire text, and a plethora of other sophisticated reasons.

Our initial experiment will not be dealing with the complexity of many types of errors but examines the simplified model described in earlier section (Section 3). Correspondingly, our numerical MCS experiment to assess parameters of this distribution is described as follows:

1. We take a sample size N = 2000 for repeated process.

2. We generate the random distribution of errors in the entire "population" of all the sentences of the material.

3. The number of errors found in these samples represent error distribution of a total number of errors in a sample.

4. We use the large number of sampled data to estimate the entire collection of materials, the error distribution, mean and confidence intervals of such distribution.

5. We take the same error density assumption of 0.07 as in the earlier section (Section 3.3).

Post-Editing Distance (PED) score remains one of the popular measurement. The PED score is often tracked on segment level, in comparison to document or system level, to examine how good the MT output is in comparison to human edited final translation.

We further conduct analysis of confidence interval (CI, or Δ) for average PED score depending on the sample size.

The absolute PED score is the minimal number of deletion and insertion operation/editing steps form initial candidate translation to post-edited translation text, divided by the length of source sentence. Because the number editing steps can be larger than the number of words in the source sentence, the absolute PED score can be greater than 1.

However, since we design to compare the PED with vector similarity, the absolute PED score needs to be normalized to a [0, 1] range. We propose a normalization function of PED (represented as PEDn) as bellow:

$$PEDn = 1 - tanh(c \cdot PED)$$

where c is a parameter defining the value of PED which brings the value of normalized PED to [0, 1], as shown in Figure 7.

Using the MCS methods we introduced in the last section (Section 4), we conduct numerical experiment to estimate confidence interval (Delta) for average normalized measurement PEDn as function of sample size (number of sentences).

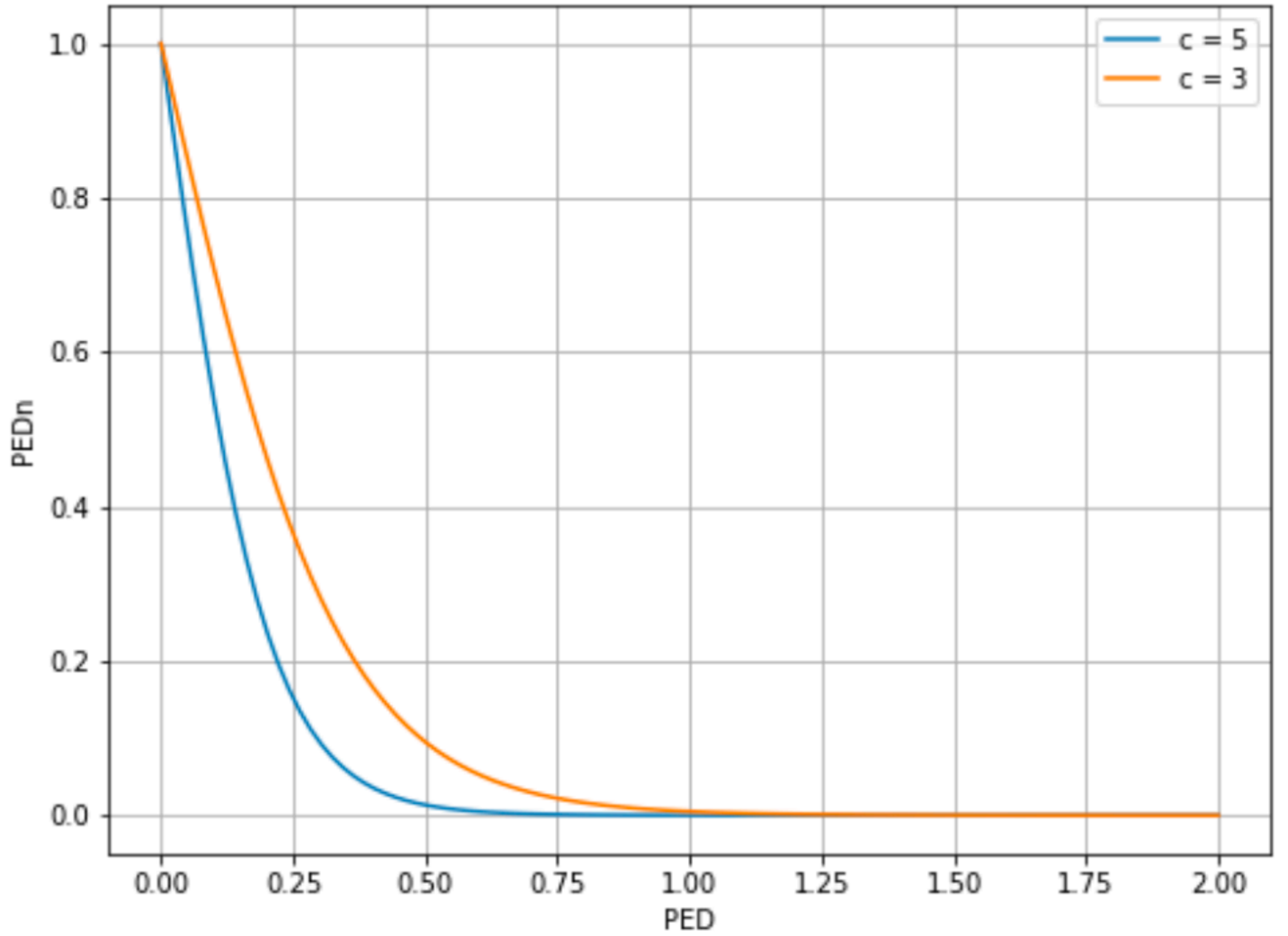


Fig. 7: Relationship between plain PED and normalized PEDn.

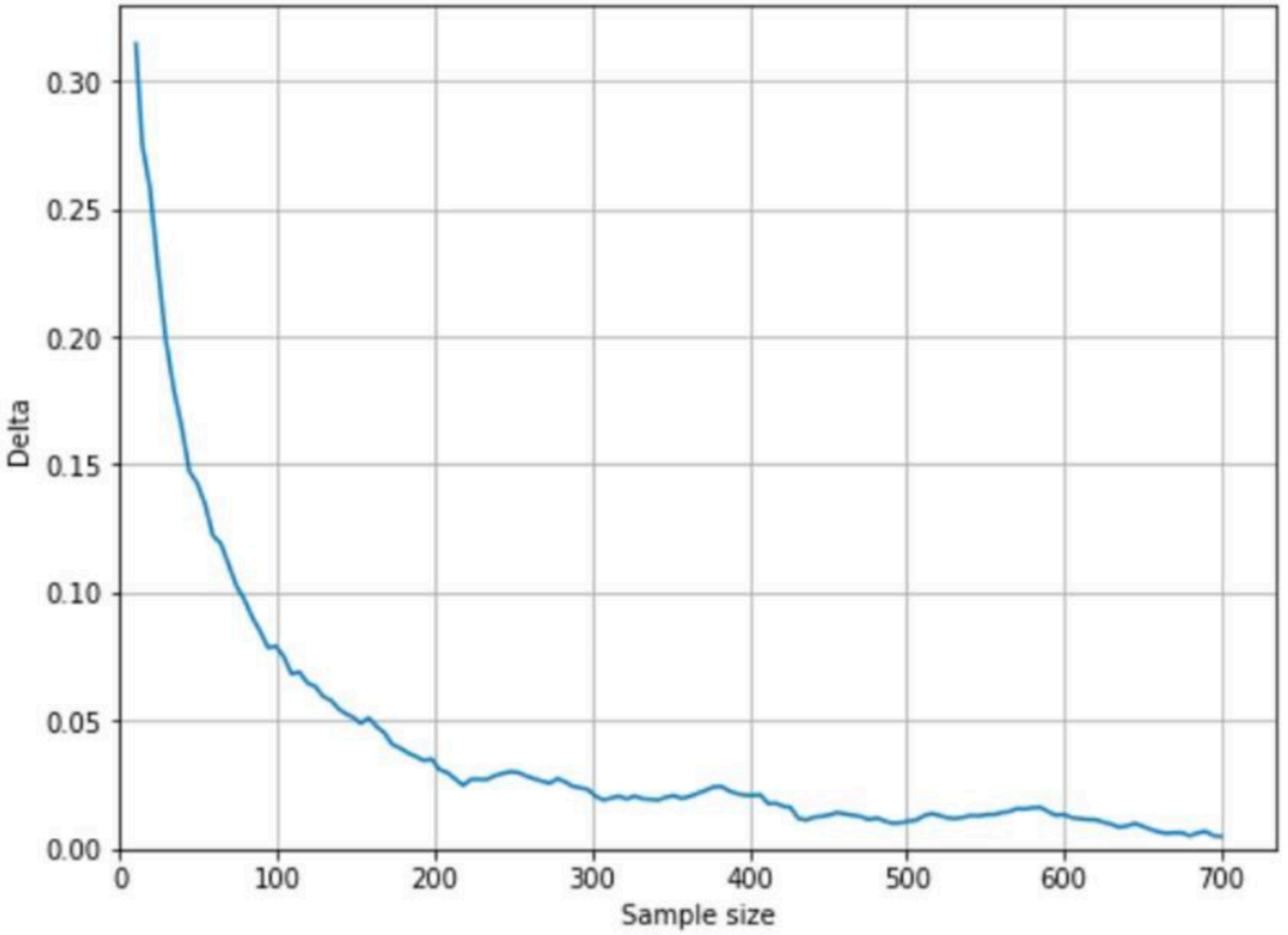


Fig. 8: Delta as function of sample size.

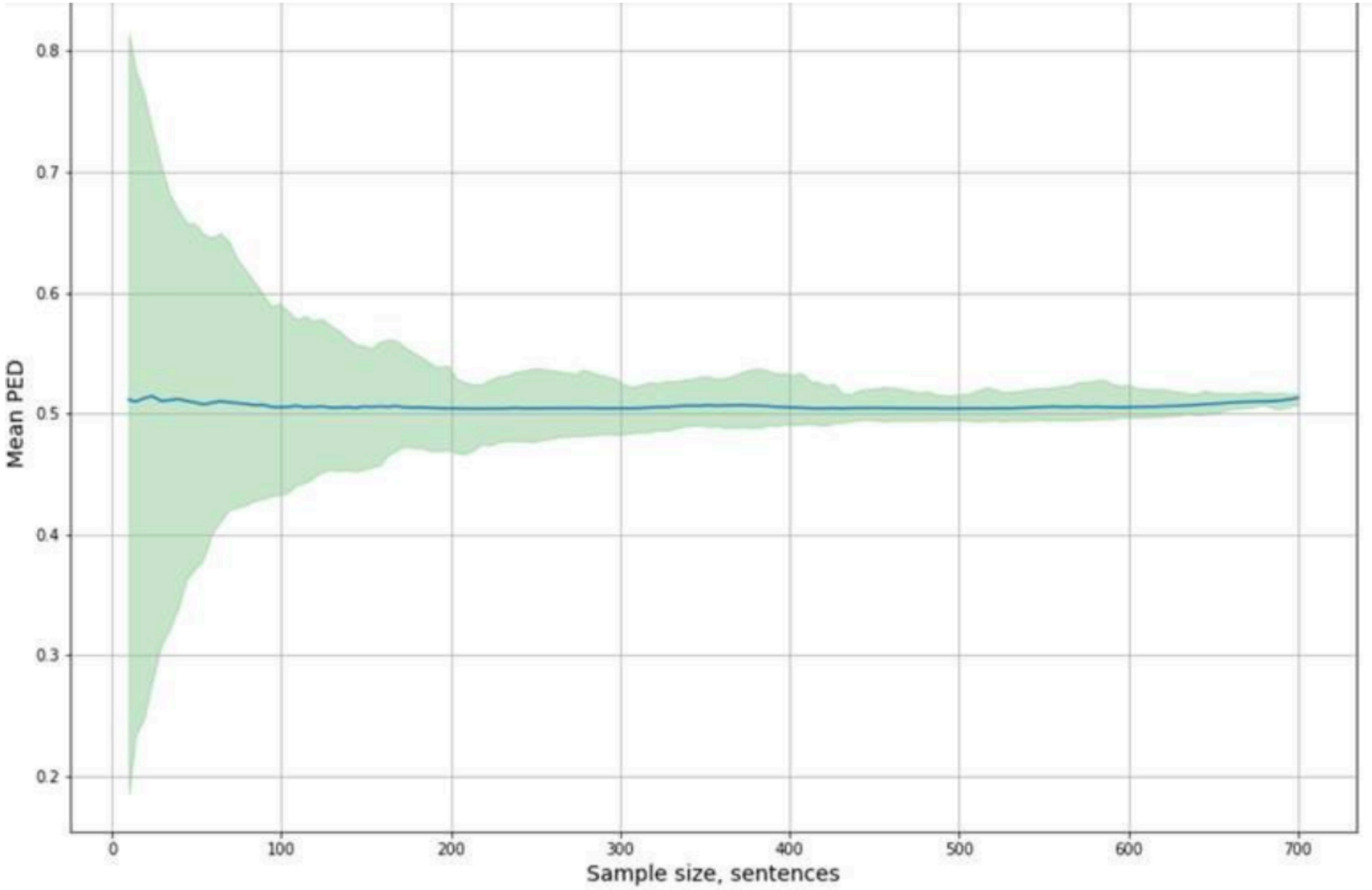


Fig. 9: A 95% confidence level credible interval for mean PEDn.

Our statistical modeling using MCS suggests that a sample size of less than 200 sentences can not reflect the overall material quality in a confident enough level on translation quality evaluation task. Using MCS, we also reduced the suggested sample size from 10k words (around 625 sentences, from Bernoulli statistics in Section 3) to 4k (Section 5) for reliable estimation of overall translation quality.

Furthermore, we suggest that, ideally, a reliability level of analytic sample quality measurement can be added to every analytic TQE scorecard in the form of confidence interval at certain confidence level as one important indicator of the level of certainty of measurement results. In the future work, we plan to compare different sampling methods, as well as apply the confidence estimation model into broader TQE metrics.

Content

- Background (& motivation of this work)
- Related work (earlier surveys)
- Our classifications of TQA
- **Discussion and Perspectives**
- Conclusions
- => Appendices (references, codes, platforms)

cartoon, https://www.bbc.com/news/

Discussion and Perspectives

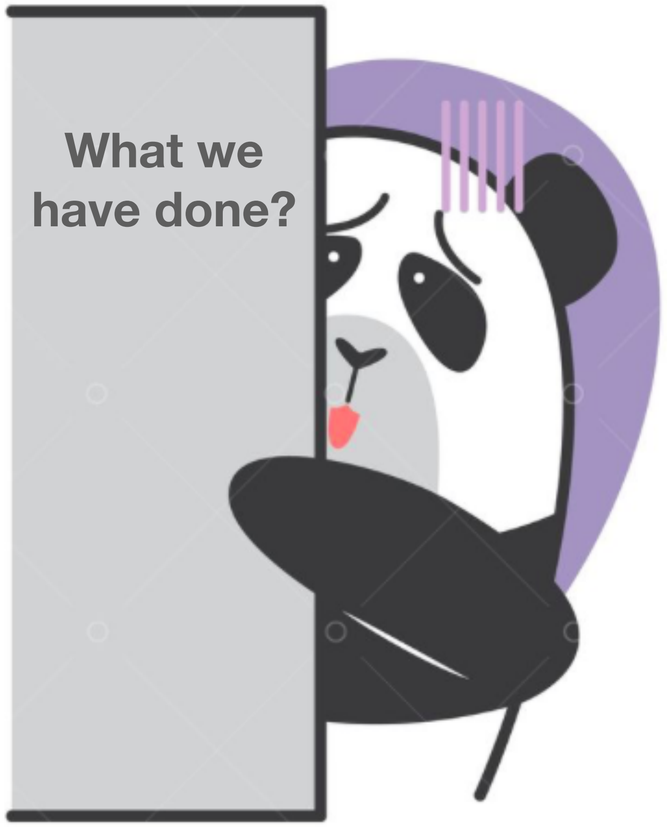
- MQM, getting more attention recently, applied to WMT tasks (QE2020), many criteria, needs to be adapted to specific task of evaluations.
- QE more attention, more realistic in practical situation when no ref available, e.g. imergence situation when MT needed, such as war, natural disaster, rescue, in some low resource scenario. As SAE stated 2001: the poor MT means leading to death, ….
- Human assessment agreement has always been an issue, and recent google-HE-with-WMT data shows the crowd source human assessment disagree largely with professional translators. i.e. the crowd source assessors cannot distinguish real improved MT system that may have syntax or semantic parts improved.
- How to improve the test set diversity (domains), reference translations coverage - semantical equivalence, etc.

Discussion and Perspectives

- Metrics still facing challenges in segment level performance/correlation.
- Metrics score interoperation: automatic evaluation metrics usually yield meaningless score, which is very test set specific and the absolute value is not informative:
- - what is the meaning of -16094 score by the MTeRater metric, or 1.98 score by ROSE?
- - similar goes to 19.07 by BEER / 28.47 by BLEU / 33.03 by METEOR for a mostly good translation
- - we suggest our own LEPOR, hLEPOR, cushLEPOR where relatively good/better MT system translations were reported scoring 0.60 to 0.80, with overall (0 ~ 1)
- Document-level Eval, Context aware MTE, gender aware
- MWEs, Literature domain MTeval, Semantics
- Large PLM evaluation model distillation

Content

- Background (& motivation of this work)
- Related work (earlier surveys)
- HumanEval, AutoEval, MetaEval
- Discussion and Perspectives
- **Conclusions**
- => Appendices (references, codes, platforms)



cartoon, https://images.app.goo.gl/Y6bAfr9oFsswWjUY7

Conclusions

- A brief overview of both human/manual and automatic assessment methods for translation quality, and the meta-assessment: assessing the assessment methods.
- Automated models cover both reference-translation dependent metrics and quality estimation without reference. There is a new trend of reference less metric.
- MT is still far from reaching human parity, especially in broader domains, topics, language pairs. TQA will still play a key role in the MT development.
- TQA/MTE methods will surely have an influence in other NLP tasks and be impacted vice versa.

Content

- Background (& motivation of this work)
- Related work (earlier surveys)
- Our classifications of TQA
- Discussion and Perspectives
- Conclusions
- **=> Appendices (references, codes, platforms)**

Selected References

- Our earlier work related to this survery:
  - Aaron Li-Feng Han and Derek F. Wong. (2016) Machine translation evaluation: A **survey** https://arxiv.org/abs/1605.04515 updated 2018 =>
  - Lifeng Han (2018) Machine Translation Evaluation Resources and Methods: A **Survey** https://arxiv.org/abs/1605.04515v8
  - Han et al. (2021) Translation Quality Assessment: A Brief Survey on Manual and Automatic Methods. https://aclanthology.org/2021.motra-1.3/
  - Lifeng Han (2022) An **Overview** on Machine Translation Evaluation. https://arxiv.org/abs/2202.11027 (in Chinese, English update forthcoming)
  - Lifeng Han, Gareth J. F. Jones, and Alan Smeaton. 2020. **MultiMWE**: Building a multi-lingual multiword expression (MWE) parallel corpora. In LREC.
  - Lifeng Han, Gareth Jones, and Alan Smeaton. 2020. **AlphaMWE**: Construction of Multilingual Parallel Corpora with MWE Annotations. In Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, pages 44–57, online. Association for Computational Linguistics.
  - Lifeng Han. (2014) **LEPOR**: An Augmented Machine Translation Evaluation Metric. Msc **thesis**. University of Macau. https://library2.um.edu.mo/etheses/b33358400_ft.pdf
  - Han (2022) PhD **Thesis** 'An investigation into multi-word expressions in machine translation' https://doras.dcu.ie/26559
  - Gladkoff and Han. 2022LREC. **HOPE**: A Task-Oriented and Human-Centric Evaluation Framework Using Professional Post-Editing Towards More Effective MT Evaluation.
  - Measuring **Uncertainty** in Translation Quality Evaluation (TQE). 2022LREC. Gladkoff, Serge ; Sorokina, Irina ; Han, Lifeng ; Alekseeva, Alexandra
  - Lifeng Han, Irina Sorokina, Gleb Erofeev, and Serge Gladkoff. 2021. **cushLEPOR**: customising hLEPOR metric using Optuna for higher agreement with human judgments or pre-trained language model LaBSE. In Proceedings of the Sixth Conference on Machine Translation, pages 1014–1023, Online. Association for Computational Linguistics.
- TQA interaction with other NLP evaluations:
  - Liu et al. Meta-evaluation of Conversational Search Evaluation Metrics, (2021) https://arxiv.org/pdf/2104.13453.pdf ACM Transactions on Information Systems
  - Pietro Liguori et al. 2021. Shellcode_IA32: A Dataset for Automatic Shellcode Generation. https://arxiv.org/abs/2104.13100
  - Why We Need New Evaluation Metrics for NLG. by Jekaterina Novikova et al 2017emnlp. https://www.aclweb.org/anthology/D17-1238/
  - Evaluation of text generation: A survey. (2020) A Celikyilmaz, E Clark, J Gao - arXiv preprint arXiv:2006.14799, - arxiv.org
  - SCOTI: Science Captioning of Terrain Images for data prioritization and local image search. D Qiu, B Rothrock, T Islam, AK Didier, VZ Sun… - Planetary and Space …, 2020 - Elsevier
- WMT findings:
  - WMT and Metrics task: Koehn and Mons (2006), …, Ondřej Bojar, et al. (2013),…on to .Barrault et al. (2020)
  - QE task findings: from (Callison-Burch et al., 2012), …on to Special et al. (2020)

Codes and Platforms

- LEPOR: https://github.com/aaronlifenghan/aaron-project-lepor
- hLEPOR: https://github.com/poethan/LEPOR (older https://github.com/aaronlifenghan/aaron-project-hlepor)
- cushLEPOR: https://github.com/poethan/cushLEPOR
- HPPR: https://github.com/aaronlifenghan/aaron-project-hppr
- MultiMWE: https://github.com/poethan/MWE4MT
- AlphaMWE: https://github.com/poethan/AlphaMWE
- HOPE: https://github.com/lHan87/HOPE
- This tutorial: https://github.com/poethan/LREC22_MetaEval_Tutorial

Acknowledgement

- We thank our colleagues who are in contribution to our work: Irina Sorokina, Gleb Erofeev.

- We thank the feedback and discussion on this tutorial structure and presentation form NLP group in The University of Manchester, especially Viktor Schlegel, Nhung Nguen, Haifa, Tharindu, Abdullah. A link to NLP at UniManchester https://www.cs.manchester.ac.uk/research/expertise/natural-language-processing/

- We also thank the funding support from Uni of Manchester (via Prof Goran Nenadic's project) and previous funding from ADAPT Research Centre, DCU. A link to NLP at UniManchester https://www.cs.manchester.ac.uk/research/expertise/natural-language-processing/ and ADAPT https://www.adaptcentre.ie/