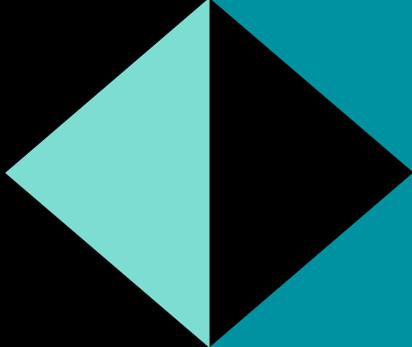# Machine translation evaluation from 'the other side of the pond'

Constantin Orasan

Centre for Translation Studies, University of Surrey
https://www.surrey.ac.uk/people/constantin-orasan

UNIVERSITY OF SURREY

While we get ready to start please open a browser and go to pollev.com/corasan432

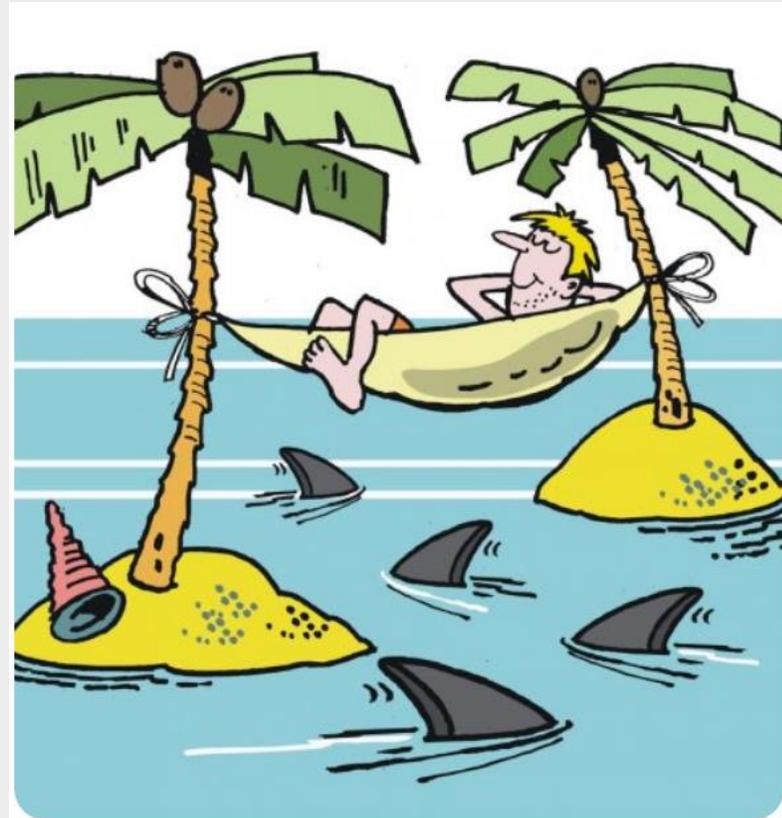# How are you today?

Please go to
pollev.com/corasan432

Why this title?

Natural Language Processing

Translation studies

Structure

1. Introduction

2. The BLEU score

3. Other evaluation metrics

4. Translation as Human Computer Interaction

5. Conclusions

AI translation advances are now posing a risk to jobs

Will Machine Learning AI Make Human Translators An Endangered Species?

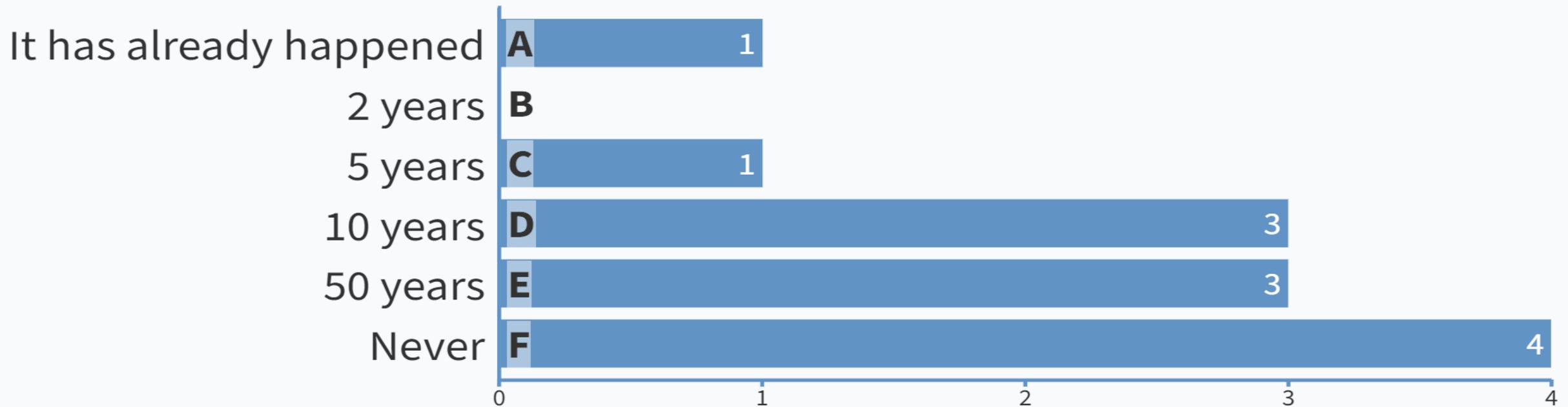Is the era of artificial speech translation upon us?

Do you believe that computers will replace human translators/interpreters in the next:

- It has already happened
- 2 years
- 5 years
- 10 years
- 50 years
- Never

Part 1: Introduction

A few definitions

Machine translation = "the field in language processing concerned with the automatic translation of texts from one (source) language into another (target) language" (Specia and Wilks, 2016)

Computer aided translation = "the use of software to assist human translators in the translation process" (https://en.wikipedia.org/wiki/Computer-assisted_translation)

Evaluation in NLP = "the process of establishing the performance of a system in a specific NLP task with respect to some predefined criteria/metrics"

Evaluation in MT

» The purpose of machine translation is to produce a translation of the source which is both fluent (**fluency**) and represents the information from the source accurately (**fidelity**).

» Fluency and fidelity are highly subjective, difficult to measure and usually depend on the context

» MT evaluation is difficult:
  • there is more than one correct translation
  • different translations can be partially correct, but in different ways

» Development of MT engines depends very much on the availability of reliable ways to assess their results

» *"The evaluation of MT is more developed than MT itself"* (Wilks 2009)

» But looking at current NLP papers some researchers know only **BLUE** 😟😭😣

How evaluation looks in NLP papers
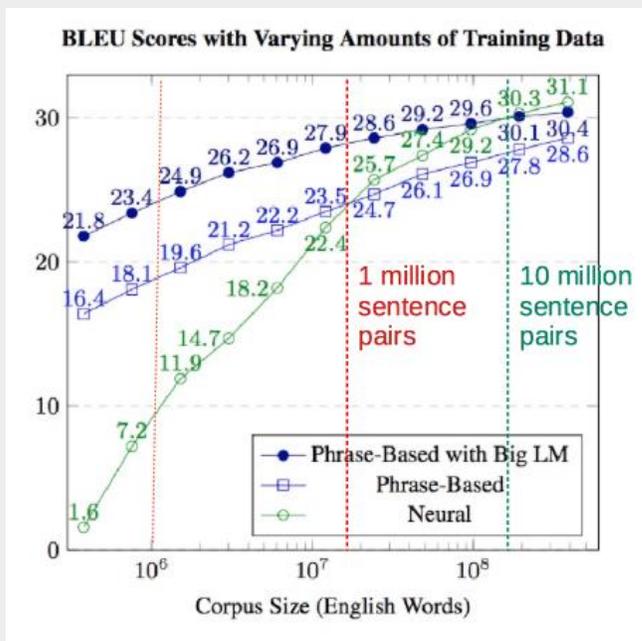
Table 4: Single model results on WMT En→Fr (newstest2014)

Model	BLEU	CPU decoding time per sentence (s)
Word	37.90	0.2226
Character	38.01	1.0530
WPM-8K	38.27	0.1919
WPM-16K	37.60	0.1874
WPM-32K	38.95	0.2118
Mixed Word/Character	38.39	0.2774
PBMT [15]	37.0	
LSTM (6 layers) [31]	31.5	
LSTM (6 layers + PosUnk) [31]	33.1	
Deep-Att [45]	37.7	
Deep-Att + PosUnk [45]	39.2	

Wu et al (2016)

Hassan et al (2018)

BLEU Scores with Varying Amounts of Training Data

Koehn and Knowles (2017)

SystemID	Settings	BLEU
Base	Transformer Baseline	24.2
BT	+Back Translation	25.57
Base8K	BT + 8K d_{ff}	26.13
CED1	Base8K + 35M CED + dropout=0.1	26.68
CED2	Base8K + 50M CED + dropout=0.1	26.61
SV1	Base8K + 35M + dropout=0.1	27.60
SV2	Base8K + 50M + dropout=0.1	27.45
SV3	Base8K + 35M + dropout=0.2	27.67
SV4	Base8K + 50M + dropout=0.2	27.49

Table 2: Evaluation Data selection results on the WMT 2017 Chinese-English test set

Part 2: The BLEU score

What is BLEU?

» BLEU = Bilingual Evaluation Understudy Score (Papineni, 2002)

» Calculates **automatically** the "similarity" between an automatic translation (hypothesis) and one or several reference translation (human translation)

» The facto evaluation metric used by MT researchers

» Shows good correlation with human judgements (… but this is challenged by some)

» Played an important role in the development of the field

- provides an easy and cheap way to repeatedly assess an MT engine
- it is language independent
- enables comparison between systems run on the same data

What is BLEU?

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert <u>after the</u> Guam <u>airport and its</u> offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as <u>the airport</u> .

Machine translation:
The American [?] international <u>airport and its</u> the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on <u>the airport</u> to start the biochemistry attack , [?] highly alerts <u>after the</u> maintenance.

- N-gram precision (score is between 0 and 1)
  - What percentage of machine n-grams can be found in the reference translation?
  - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the the")
  - The values for BLEU are rather low: a human translator scored 0.3468 against four references and 0.2571 against two references. (Papineni, 2002)
  - Usually N=4

- Brevity penalty
  - Can't just type out single word "the" (precision 1.0!)

- Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

Slide adapted from Bonnie Dorr

What is BLEU?

Reference translation 1:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:
US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

Slide from Bonnie Dorr

Rank translations

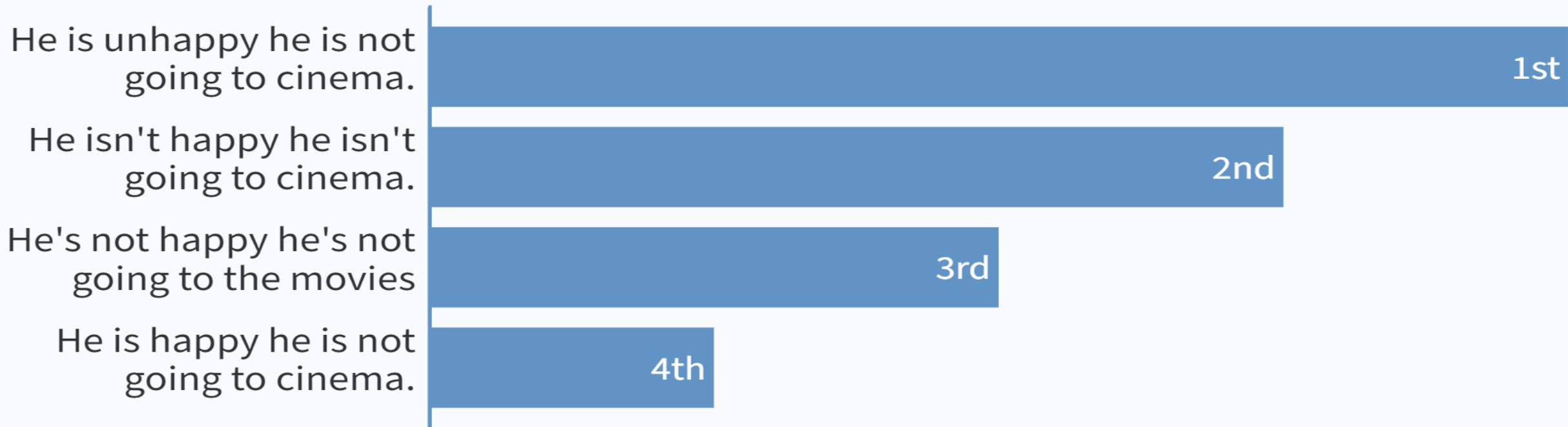» Source: Nu e bucuros că nu merge la cinema

» MT outputs (not real outputs)

1. He is unhappy he is not going to cinema.

2. He is happy he is not going to cinema.

3. He isn't happy he isn't going to cinema.

4. He's not happy he's not going to the movies

Now let's see what the BLUE score tells us: https://www.letsmt.eu/Bleu.aspx

Please go to pollev.com/corasan432

Answer to question

Rank the translation of "Nu e bucuros că nu merge la cinema"

He is unhappy he is not going to cinema. — 1st

He isn't happy he isn't going to cinema. — 2nd

He's not happy he's not going to the movies — 3rd

He is happy he is not going to cinema. — 4th

Powered by Poll Everywhere

What BLEU thinks?

Source: Nu e bucuros că nu merge la cinema

Reference: He is not happy he is not going to cinema.

1. He is unhappy he is not going to cinema. (63.98)
2. He is happy he is not going to cinema. (75.17)
3. He isn't happy he isn't going to cinema. (23.90)
4. He's not happy he's not going to the movies (20.10)

We use the implementation from https://www.letsmt.eu/Bleu.aspx

Problems with BLEU

» Does not consider the meaning, but unigrams capture *fidelity*

» Does not consider grammar, but 3-grams & 4-grams account for *fluency*

» Mismatch of a function word is treated the same as a difference in a content words

» Sensitive to morphological variations

» Sensitive to tokenisation

» What does a 0.32123 BLEU score mean?

The purpose of machine translation is to produce a translation of the source which is both fluent (**fluency**) and represents the information from the source accurately (**fidelity**).

To wrap up about BLEU

» Positives

- is a good measure when evaluating at corpus level

- is fast and fairly well understood by the community

- used in many "generation-like" tasks

» Negatives

- fiddly at sentence level (but there are variants for sentence BLEU)

- not good for specific problems (e.g. translation of sentences that contain sentiments/emotions (Saadany and Orasan, 2020))

- does not tell you what is wrong with a translation

» … but there are other ways to evaluate

Part 3: Moving towards the pond ☺

Evaluation in MT

» It is a very well research field

» Some of the methods provide more information than others

» Not all the errors are equal:

- "Vă mulțumesc foarte mult, doamnă președintă și membri oribili." (automatic translation from an EU speech) (ASR source: "Thank you very much, Madam President and horrible members.")
- Errors in translating healthcare documents can be dangerous

» When performing MT evaluation we need to decide:

- how the system is considered
- who performs the evaluation
- what is assessed

How the system is considered

Glass box

» We assess the contribution of different components to the final translation

» Quite difficult because sometimes it is not obvious the source of an error

» Challenging to use with data-driven MT methods

» Could involve checking whether untranslated words are not present in the translation tables or embeddings

» Could be used in some speech-to-speech translation systems

Black box

» The system is seen as a black box with input(s) and output(s)

» Most common way to evaluate MT systems (and automatic systems in general)

» Usually easier to apply

» Does not give many insights why an error occurs

» BLEU!!!!

How the evaluation is performed

Manual (human) evaluation

- involves humans reading and assessing the translation (could be done using crowdsourcing) according to some guidelines

- theoretically, it is the most reliable type of evaluation

- in reality is expensive, slow and possibly unreliable

- cannot be used when developing automatic systems

Examples

- assessment of fluency and fidelity on a scale

- pairwise comparison or ranking (*which one is better?*)

- error analysis

Automatic evaluation

- assumes there is a way to automatically determine quality

- usually requires "reference" translation(s)

- focuses largely on fidelity (the assumption being that it will indirectly measure fluency as well)

- numerous metrics proposed which calculate the overlap between the automatic translation and the gold translations: BLUE, METEOR, NIST, WER …

- quality estimation does not require a reference translation

- some participants in shared tasks "gamed the metrics"

What do they mean? Do they correlate with human judgements?

What is assessed

Intrinsic evaluation

- we evaluate the output directly

- most of the discussion so far

- may be artificial and not too meaningful (*what does a 0.42 BLEU score mean for a user?*)

- could be good for comparing systems, but

- may not capture the use of the translations

Extrinsic evaluation

- we use the translation in a specific task and we measure the performance on that task

- usually focuses on measuring the presence of the information from the source, but may not measure the "quality" of a translation

- usually cannot be automated and can be quite difficult to implement

Examples

- Cloze tests, reading comprehension tests

- postediting effort, complete a task

Part 4: Translation as human computer interaction

What is HCI

» Human-Computer Interaction (HCI) = the study of interaction between people, computers and tasks

» HCI is a multidisciplinary field which is not only about user interface. It considers:

- human factors: how people interact with the tools
- ergonomics: the ease hardware (and recently software) can be used

» In the context of translation and interpreting **cognitive ergonomics** – *the cognitive demands placed on users by the design and complexity of computer programs* – is particularly important

Translation as HCI

- *"translation as human-computer interaction is one in which varying levels of repetition are characteristic, making the task suitable for translation memory tools. High volume is also a typical feature, as is the need to complete the translation task under significant time pressure, making Machine Translation a potentially suitable translation aid."* (O'Brian 2012)

- collaborative (volunteer) translation or subtitling and dubbing of audio-visual material also involve interaction with computers

- even literary translation may involve interaction with computers (use of dictionaries, corpora) and can be seen as HCI

» The last 30 years have seen a dramatic change in the way technology is used in translation which in turn had a dramatic impact on the translation profession. Advances of Neural Machine Translation in the last 5 years.

» This brought benefits, but also challenges

Project management triangle

» aka *triple constraint* or *iron triangle*

» In a translation project we have:

- Time: time available to deliver the project

- Cost: the amount of money available

- Quality: fit-for-purpose that the translation must achieve to be considered successful

» In a translation project what it the most important aspect: time, cost or quality. Vote now at **pollev.com/corasan432**

» Most clients would like a translation job to be done quickly, cheap and well. ☺

» but no side can be altered without changing the other ones (i.e. these constraints are in competition)
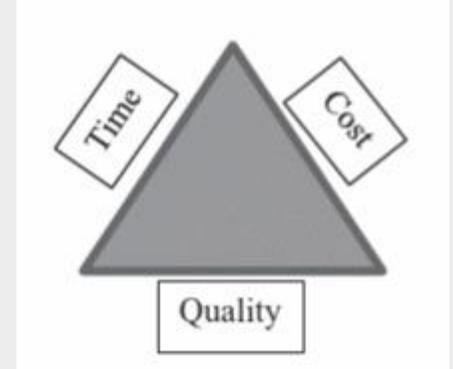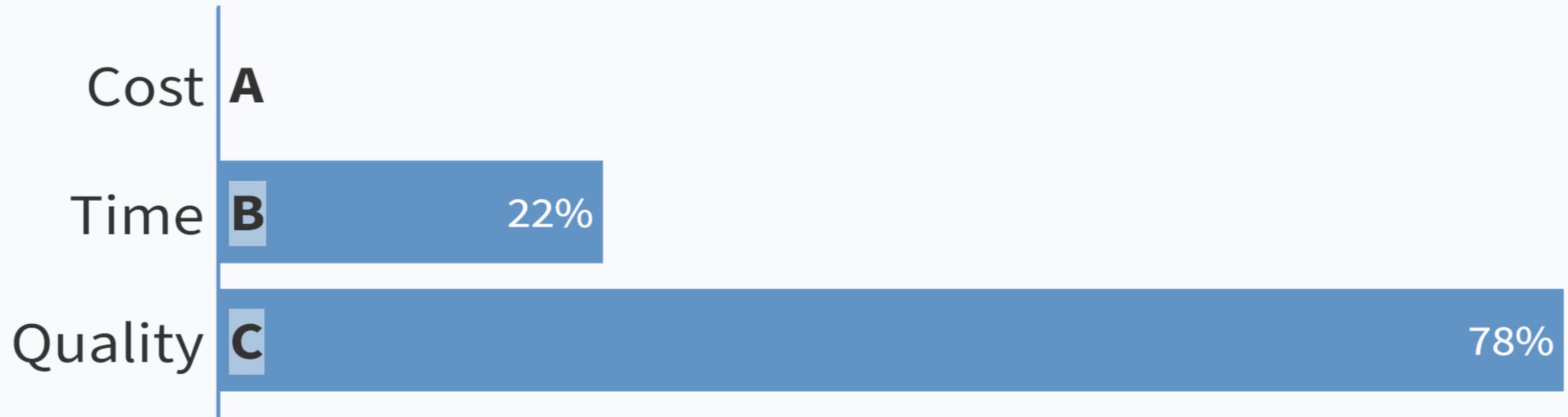
Benefits of using technology

- Use of technology increases the **productivity** of translators

- Increases the **quality** of translation

- Reduces the **costs** for producing content

- these statements are particularly true when talking about translation memories: less typing, more consistent, less effort

- they are still being discussed when using machine translation: much more content can be translated, but is it good enough?

Challenges

- The technology introduced some significant changes to how translators work (ages ago some were even against using a word processor)

- Translators may feel dehumanised by the technology they are required to use (this is particularly in postediting tasks which is a less creative task, the pay rate goes down and the expectations in terms of productivity goes up)

- MT is a black box and translators have limited influence on its performance/behaviour (the system does not incorporate the feedback from translators)

- Recycling previously translated segments leads to 'the tendency of translated text to gravitate towards the centre of a continuum' (Baker 1996)

- Translators may not longer work with full texts

- Translators may not be aware of the help technology brings

- Question no longer "how do we know when a translation is good" but "**how do we know when a translation is good enough**"? **fit-for-purpose**

Human parity

» There have been recent claims that MT reached human parity (i.e. MT output is as good as a professional translation).

» When we try to answer this we need to keep in mind: the nature of start text, the MT system, the human translation entering into the comparison, the language pair, the definition of 'quality' and the human or automatic metrics used.

» MT is usually assessed at the segment level, whilst we should probably done at document level

» See a discussion in (Laubli, Snrich and Volk, 2018)

We may want to rephrase this for translators/clients as **cost-beneficial** in terms of effort and quality

When is appropriate to use post-editing?

» Postediting = "the process of improving a machine generated translation with a **minimum manual labour**" (TAUS 2010)

» Moorkens and Way (2016) compared the usability of TM matches and input from MT

» The results show that low-quality MT matches are not useful to the translators in over 36% of cases

» The translators described these suggestions as *irritating*

» The conclusion of the article: "*MT confidence measures need to be developed as a matter of urgency, which can be used by post-editors to wrest control over what MT outputs they wish to see, and perhaps more importantly still, which ones should be withheld*"

Quality estimation

» A research field that attempts to establish the quality of a translation without the need of a reference translation

» Could be the solution for post-editing by indicating when it takes more effort to postedit a sentence than to translate from scratch

» We tried to find out whether the quality of a translation, as determined by a QE system, influence the work of professional translators? (Parra, Bechara and Orasan, 2017)

» We train it to predict the **target-side Fuzzy Match Score** (FMS) of the translation

» FMS > 75%: consider a segment good enough to be post-edited

User study

» Uses a modified version of PET

» We used a traffic light system:
  - Light yellow: translate from scratch (Label: Translate)
  - Light blue: MT is provided but no QE information (Label: Postedit)
  - Light green: MT is provided and the translator encouraged to post-edit (Label: QE postedit)
  - Light red : MT is provided but the translator encouraged to translate from scratch (Label: QE translate)
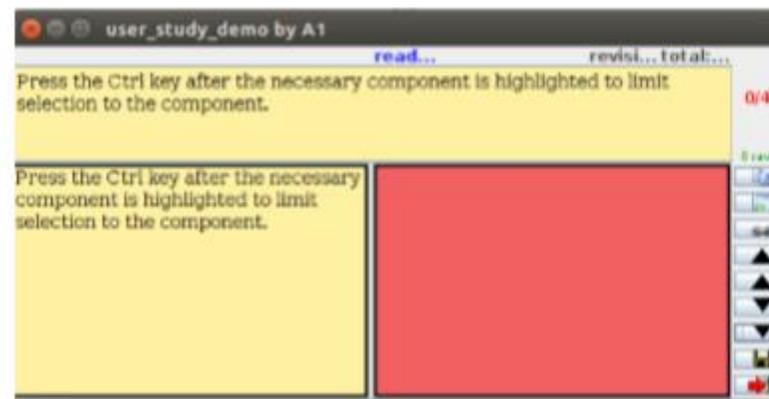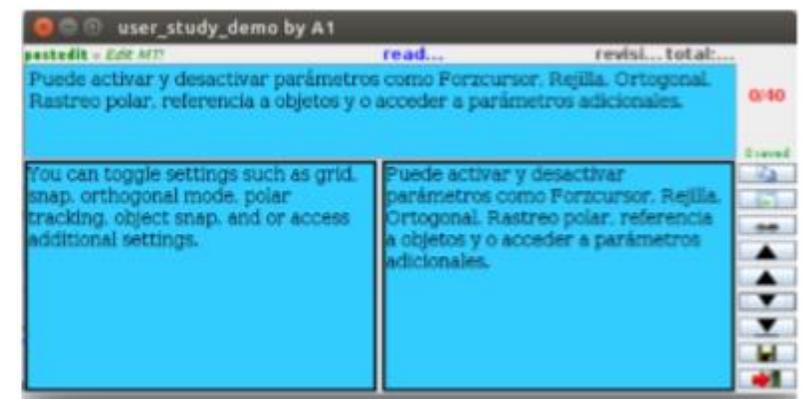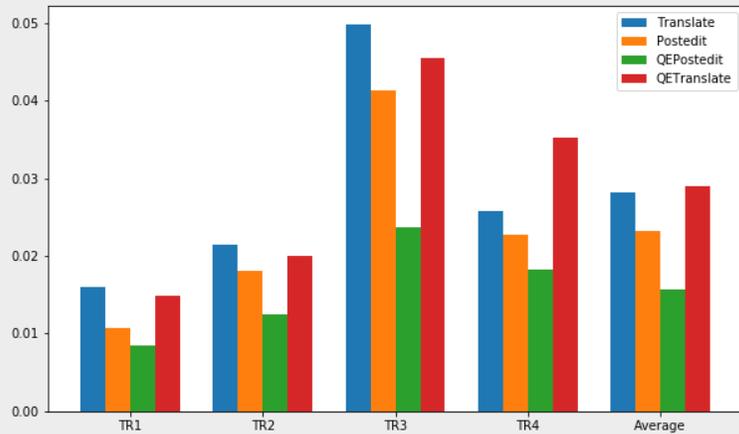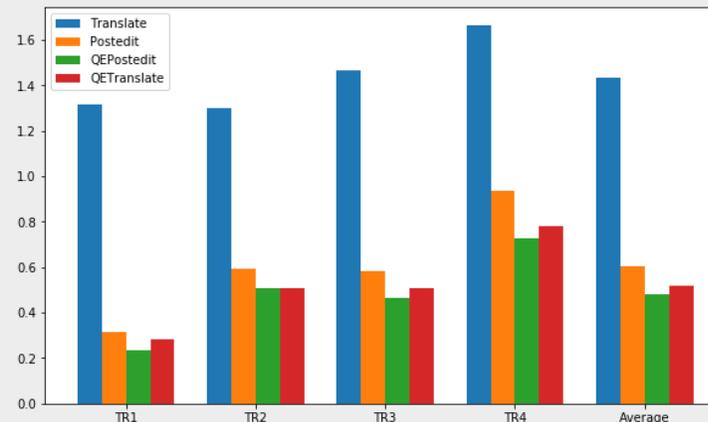

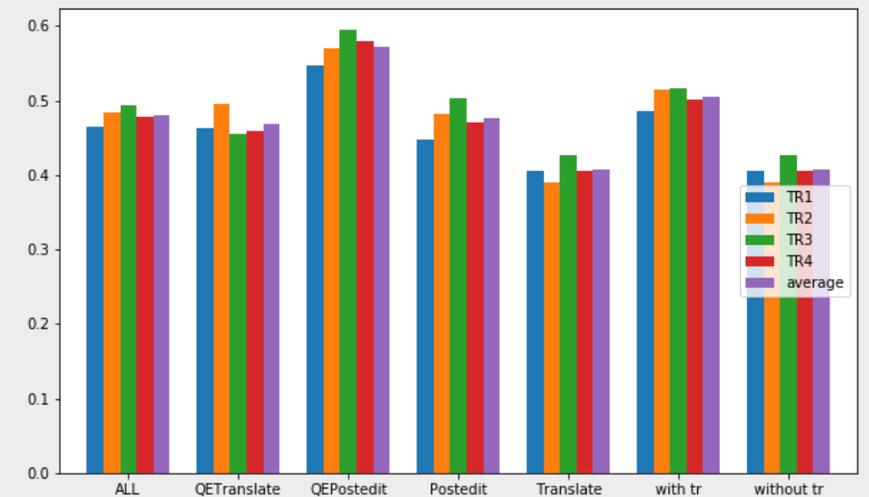Translate from scratch   Post-edit without MTQE

The productivity of translators

Normalised average seconds per word

Normalised average number of keystrokes

How different are the translations?

Conclusions of the study

» QE information can help:

- Asking translators to post-edit MT output when the estimated quality is high helps their productivity

- Showing a translation when we know it is bad does not necessary hurt, but it does not help either

- Even if the translators do not admit it, QE information helps them

» Overall providing MT can help translators

» Open questions:

- What is the threshold that we should use for suggesting post-editing

- How the performance of QE influences our findings

- Understand better the influence of showing bad translations to translations

- PET records a lot of information about keys pressed. Understand better the translators' behaviour

Conclusions

» BLEU (and its variants) are here to stay

» BLEU is very useful for some situations, but be aware of its limitations

» Machine translation is becoming more important for the translation profession

» When we evaluate MT we need to keep in mind how it is used

» When designing evaluation methods NLP researchers should work with the users (translation professionals)

Thank you!

References

» Hassan, Hany et al. (2018) "Achieving Human Parity on Automatic Chinese to English News Translation." *ArXiv* abs/1803.05567

» Philipp Koehn and Rebecca Knowles (2017) Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, Canada, August 4

» Samuel Laubli and Spence Green (2020) Translation technology research and human-computer interaction (HCI). In O'Hagan, M. (ed.) *The Routledge handbook of translation and technology*. London; Routledge.

» Sharon O'Brien (2012). Translation as human–computer interaction. *Translation SpacesTranslation Spaces A Multidisciplinary, Multimedia, and Multilingual Journal of Translation*, *1*, 101–122. https://doi.org/10.1075/ts.1.05obr

» Specia, L., & Wilks, Y. (2016). Machine Translation. In R. Mitkov (Ed.), *The Oxford Handbook of Computational Linguistics 2nd edition*. Oxford University Press.

» Saadany, H., & Orăsan, C. (2020). Is it great or terrible? Preserving sentiment in neural machine translation of Arabic reviews. *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, 24–37. https://www.aclweb.org/anthology/2020.wanlp-1.3/

» Carla Parra Escartín, Hanna Béchara, Constantin Orăsan (2017) Questing for Quality Estimation A User Study, The Prague Bulletin of Mathematical Linguistics 108, p. 343–354, https://ufal.mff.cuni.cz/pbml/108/art-bechara-escartin-orasan.pdf

» Wu, Y. et al. (2016) "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation." *ArXiv* abs/1609.08144 (2016)