Ranking MT Systems via Post-Editing

Wilker Aziz†, *Ruslan Mitkov*† *and Lucia Specia*§

† Research Group in Computational Linguistics
University of Wolverhampton
{w.aziz, r.mitkov}@wlv.ac.uk

§ Department of Computer Science, University of Sheffield
l.specia@sheffield.ac.uk

MT evaluation

Automatic → shallow, hardly accounts for variability

Manual (scoring/ranking) → expensive, subjective (agreement issues)

Post-editing

- a more natural and objective task
- can be assessed objectively in terms of edit distance, keystroke count, post-editing time, etc.
- produces translations that suit a purpose
- reaching agreement requires intensive training

Post-editing + Edit distance

Edit operations typically do not capture any notion of effort.

S	Apple **prosecuted** for patent violation.	Edits
R	Apple fue procesado por violación de patentes.	
MT_1	Apple procesado por violación de patentes.	2
MT_2	Apple **prosecuted** por violación de patentes.	2

- **same** edit distance
- MT_2 left "prosecuted" **untranslated**
- MT_1 **only missed the auxiliary** "fue"

Idea

Rely on **more informed effort indicators**, such as keystroke count and post-editing time while **coping with low inter-annotator agreement**.

- **drop** the assumption that post-editing effort indicators can be **directly compared across annotators**
- **learn how to compare** indicators produced by different annotators **from data**

Task

Ranking MT systems using post-editing effort indicators.

S	because the poll has been taken regularly since 1995	T	K	H
MT_1	porque la encuesta ha sido tomado regularmente desde 1995			
PE_{A_1}	porque la encuesta se realiza periódicamente desde 1995	20	34	0.5
MT_2	dado el votación ha sido tomado regularmente desde 1995			
PE_{A_2}	dado que la encuesta ha sido tomada regularmente desde 1995	40	14	0.36

Time, **K**eystrokes and **H**TER

- alternative translations must be post-edited by different people (avoid bias)
- people work at different paces (reading/typing speed) and approach similar tasks differently (editing strategies)
- it is hard to use **T**, **K** or (even) **H** directly to find out which task required the least effort: **H** and **K** suggest MT_1 was worse (in reality the changes were mostly stylistic), **T** suggests MT_2 was worse (but what if A_2 is a consistently slower editor?).

Approach

Build a gold-standard

Observe different people performing the same tasks → training data

S	Apple prosecuted for patent violation.
MT_i	Apple procesado por violación de patentes.

Who	**T**ime	**K**eystrokes	**H**TER
A_1	10	30	0.1
A_2	12	27	0.2
...	...	...	...
A_n	20	30	0.1
μ	15	28	0.15

- choose a notion of effort (e.g. post-editing time, keystroke count, HTER)
- assume that the gold-standard for each task is the average (considering a reasonable number of annotators)

Learn regressors

A regressor per annotator

training points are PE tasks

- map features (e.g. anything that can be extracted from source, MT and PE, as well as metadata, such as post-editing time and keystroke count) onto the "mean annotator" (gold-standard)

Experiments

AI ("as is") compare effort indicator directly

MN compare effort indicator after mean normalisation

R compare the predicted mean (using SVR)

- $\mathbf{R}_1$ only the indicator of interest is used as feature, e.g. A_1's post-editing time is the sole feature used to predict the mean post-editing time.
- $\mathbf{R}_4$ 4 indicators are used as features, e.g. A_1's PE time, keystroke count, HTER and length of the source are features in predicting the mean post-editing time.

Rank correlation

Our dataset is a subset of the manual annotation gathered as part of the WMT10's shared task. We computed rank correlation with human rankings.

- Boldface shows the best in a row
- A star shows the best in a column
- A double star show the best of all

Segment level

Target	AI	MN	R_1	R_4
time	0.0975	0.1555	0.2054	**0.2451**
keystrokes	0.1941⋆	0.2189	**0.3065**⋆	0.2870
HTER	0.1794	0.2637⋆	0.2693	**0.3559**⋆⋆

- **AI** is the worst strategy for all objectives
- Even **HTER** can be improved by our methods. Traditionally **HTER** is assumed to be comparable across editors, we show this is not always the case.

System level

Target	AI	MN	R_1	R_4
time	0.3696	**0.7333**⋆⋆	0.6969⋆	0.5757⋆
keystrokes	0.4787	**0.6121**	0.5878	0.5636
HTER	0.5393⋆	0.3939	0.4181	**0.5636**

- **AI** is the worst strategy for all objectives. If nothing is to be done, then **HTER** is indeed the most "directly comparable" indicator.
- Again, dropping the assumption that effort indicators are directly comparable across editors improves even **HTER**.
- Post-editing time performs really well to rank systems globally (as opposed to in a segment basis).

Example

Overall ranking using different objectives in their best performing setup

Human ranking	Time	HTER	Keystrokes
0.65: On-B	0.66: On-B	0.67: On-B	0.63: On-B
0.58: RBMT-3	0.54: On-A	0.56: UEDIN	0.55: On-A
0.56: On-A	0.53: UEDIN	0.55: On-A	0.53: UEDIN
0.55: PROMT	0.51: PROMT	0.53: UPC	0.52: RBMT-1
0.52: UPC	0.50: UPC	0.49: PROMT	0.51: UPC
0.52: UEDIN	0.48: RBMT-3	0.48: RBMT-1	0.49: PROMT
0.46: RBMT-4	0.46: On-C	0.46: JHU	0.45: RBMT-3
0.45: RBMT-1	0.45: JHU	0.42: RBMT-3	0.45: JHU
0.43: On-C	0.43: RBMT-1	0.41: RBMT-4	0.43: RBMT-4
0.36: JHU	0.43: RBMT-4	0.41: On-C	0.42: On-C