



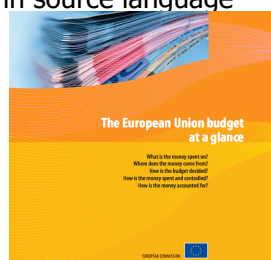
Project Scope

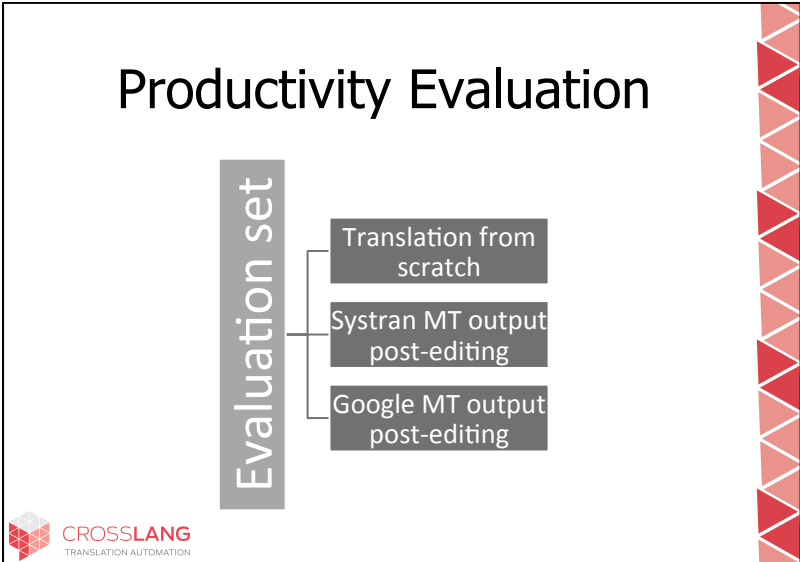
- MT Evaluation project
 - Productivity evaluation
 - Quality evaluation of the final translation
- MT technology
 - Baseline engines: Systran & Google
 - Language direction: English to Dutch
- 10 informants
 - Dutch native speakers
 - Post-graduate students Thomas More (KULeuven)
 - European Master in Specialised Translation (EMST)



Project Scope

- Evaluation set
 - Content type: European Commission publication
 - Domain: financial
 - Volume: 3246 words in source language





Productivity Evaluation

Home

Productivity Task

Source (English)
Association for Road Safety - Conference

Target (French)
Association pour la sécurité routière - Conférence

Segment: 1 of 8
Filename: Sample.doc_1211

Pause Next

- Time tracked in the background
- Throughput calculated in #words/hour for each activity



Quality Evaluation

Home

Quality Task

Source (English)
Association for Road Safety - Conference

Target (French)
Association pour la sécurité routière - conférence

Translation Quality
 Excellent Good Fair Poor

Comments

Segment: 1 of 8
Filename: Sample.doc_1211

Pause Next

- Human assessment rating individual segments 1-5 scale
- "Blind" test: "how" the translation was produced (with or without the help of MT) not disclosed to informants

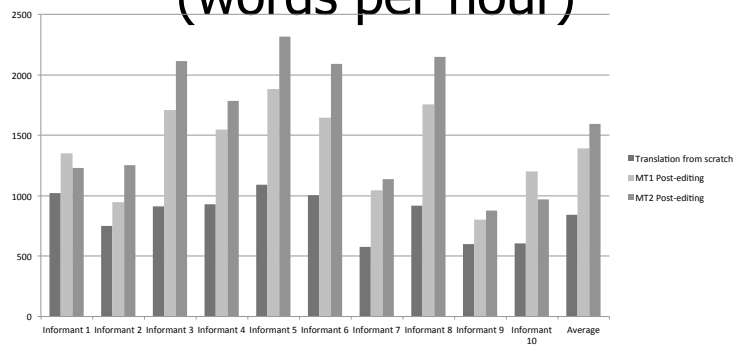


Questions asked

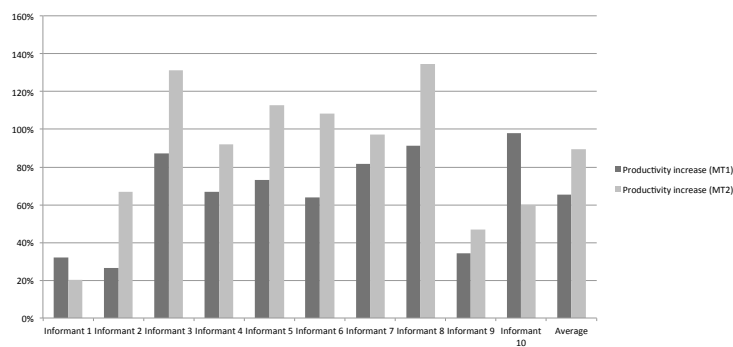
- ◆ Does post-editing MT output go faster than translating from scratch?
- ◆ If so, which MT engine scores best and how much are the respective productivity increases?
- ◆ If so, is the quality level similar/equal to the “manually” produced translations?



Average Throughput (words per hour)

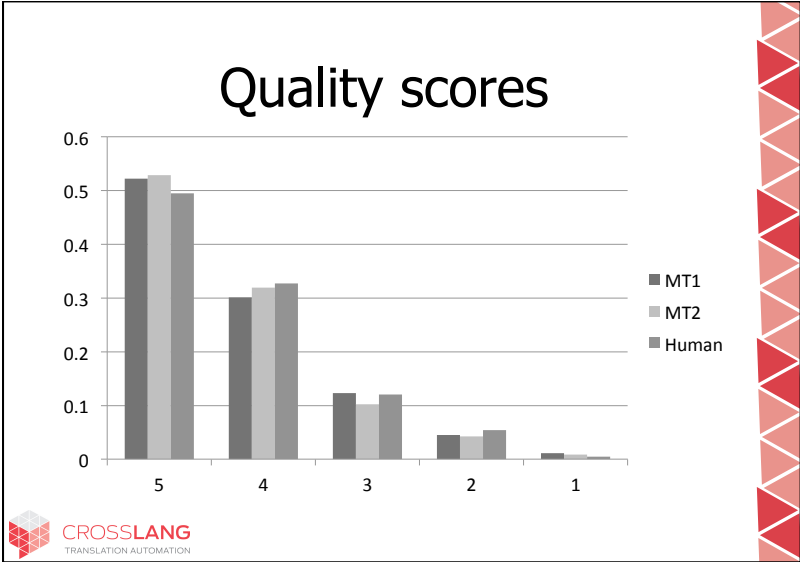


Average Productivity Increase



(Preliminary) findings

- Productivity increase in 100% of the cases
- Ranging from 20% to 134%
- Average 65% for MT1 and 89% for MT2
- Two informants worked faster with MT1, all others gained most productivity increase from working with MT2




Does quality PE MT resemble human translation?

How "bad" is PE MT?

Chi Square goodness of fit test.

Null hypothesis H0: quality assessment of PE MT is similar to quality assessment of human translations (we expect no outliers for the different categories), hence distribution of quality scores is similar.

Result: how much "better" or "worse" is PE MT in each category?


CROSSLANG
TRANSLATION AUTOMATION

Findings

94.43% is acceptable (vs. 94.16%)

	Extra %	O (MT1)	E	Proportion (MT1)	O(Human)	Proportion (Human)	(O-E)^2 / E	
More=better	5.2	5	328	310.9319	0.52146264	305	0.49432739	0.9369223
More=better	-8.96	4	189	205.9287	0.30047695	202	0.3273906	1.39164899
	2.03	3	77	75.43922	0.12241653	74	0.11993517	0.03229126
Less=better	-20.15	2	28	33.64182	0.0445151	33	0.0534846	0.9461463
Less=better	56.31	1	7	3.058347	0.01112878	3	0.00486224	5.08007445
		Total	629	629	1	617	1	8.38708331
								p= 0.07838476



Findings

94.92% is acceptable (vs. 94.16%)

	Extra %	O (MT1)	E	Proportion (MT2)	O(Human)	Proportion (Human)	(O-E)^2 / E	
More=better	6.48	5	333	311.4263	0.52857143	305	0.49432739	1.49449964
More=better	-2.61	4	201	206.2561	0.31904762	202	0.3273906	0.13394201
	-18.06	3	64	75.55916	0.1015873	74	0.11993517	1.76833782
Less=better	-24.8	2	27	33.6953	0.04285714	33	0.0534846	1.33036477
Less=better	38.74	1	5	3.063209	0.00793651	3	0.00486224	1.22458474
		Total	630	630	1	617	1	5.95172898
								p= 0.2027823







Conclusions

Given the **outspoken productivity increase** from post-editing MT output as compared to translating from scratch, we can conclude that **final translation quality** is not affected by using MT – on the contrary, a (slight) overall quality increase has been observed.

It is important to note that the contribution of all categories to Chi Square is fairly balanced, except for category 1 (MT1). On a very large scale this may become problematic.



Future work

- Include scores for original MT output
- Include inter-annotator agreement
- Investigate whether category 1 can be predicted and whether corrective measures can be taken accordingly
- Include metrics for PE effort
- Include control set

- Aim: consolidate our current findings and measure impact of PE on MT output
- Aim: investigate impact of correcting outliers in category 1