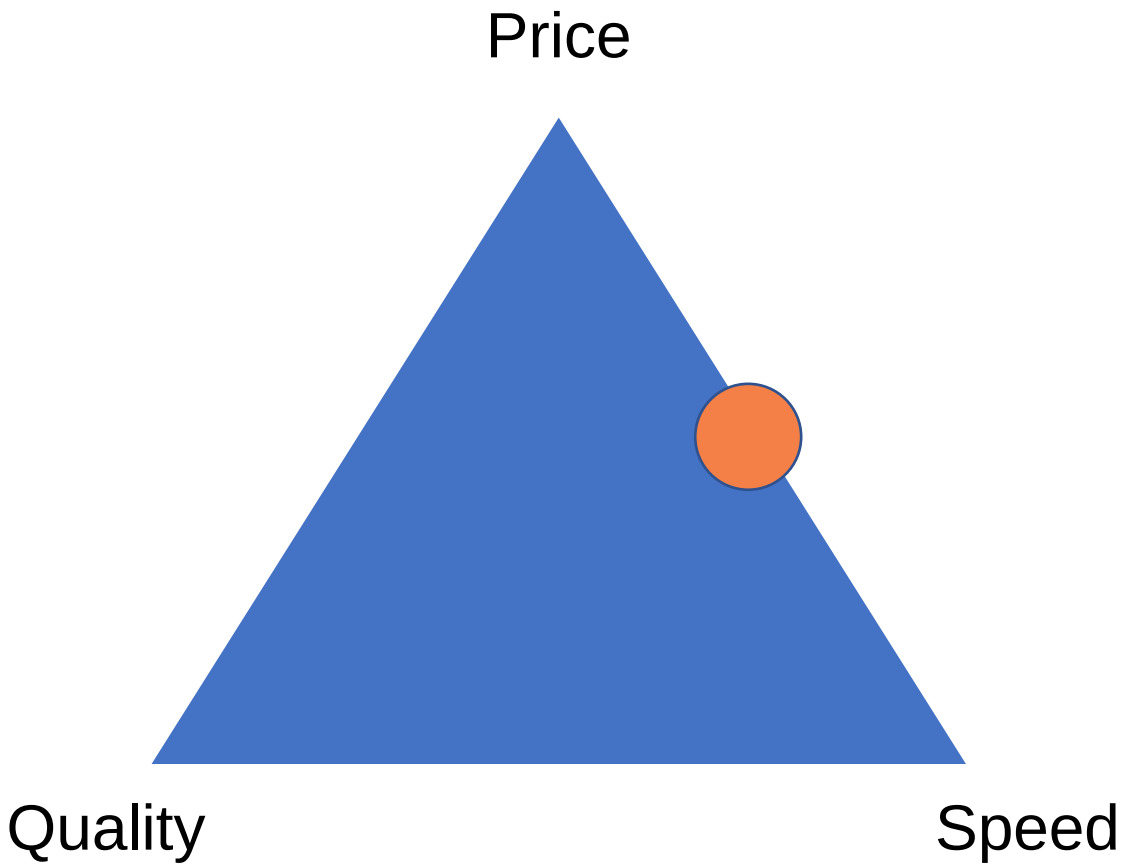
Factor 1000: Better MT, more content, and what we can do with it

Chris Wendt
Group Program Manager – Microsoft
chris.wendt@microsoft.com
@Tian500

The only job:

Remove the language barrier

Why MT?

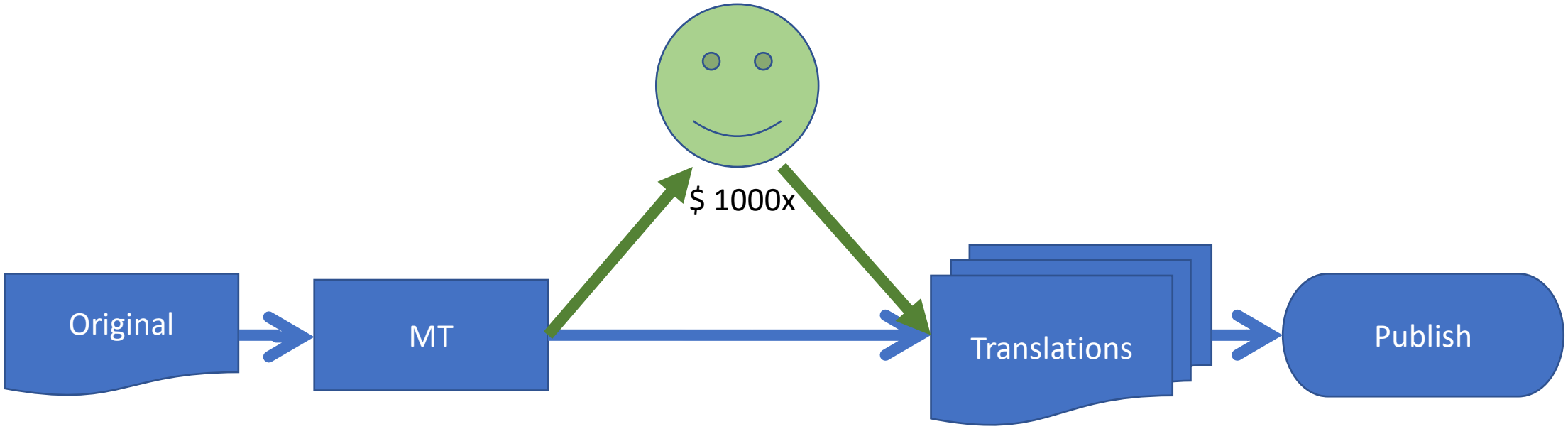


	Human	Machine
1M words	$250000	$280

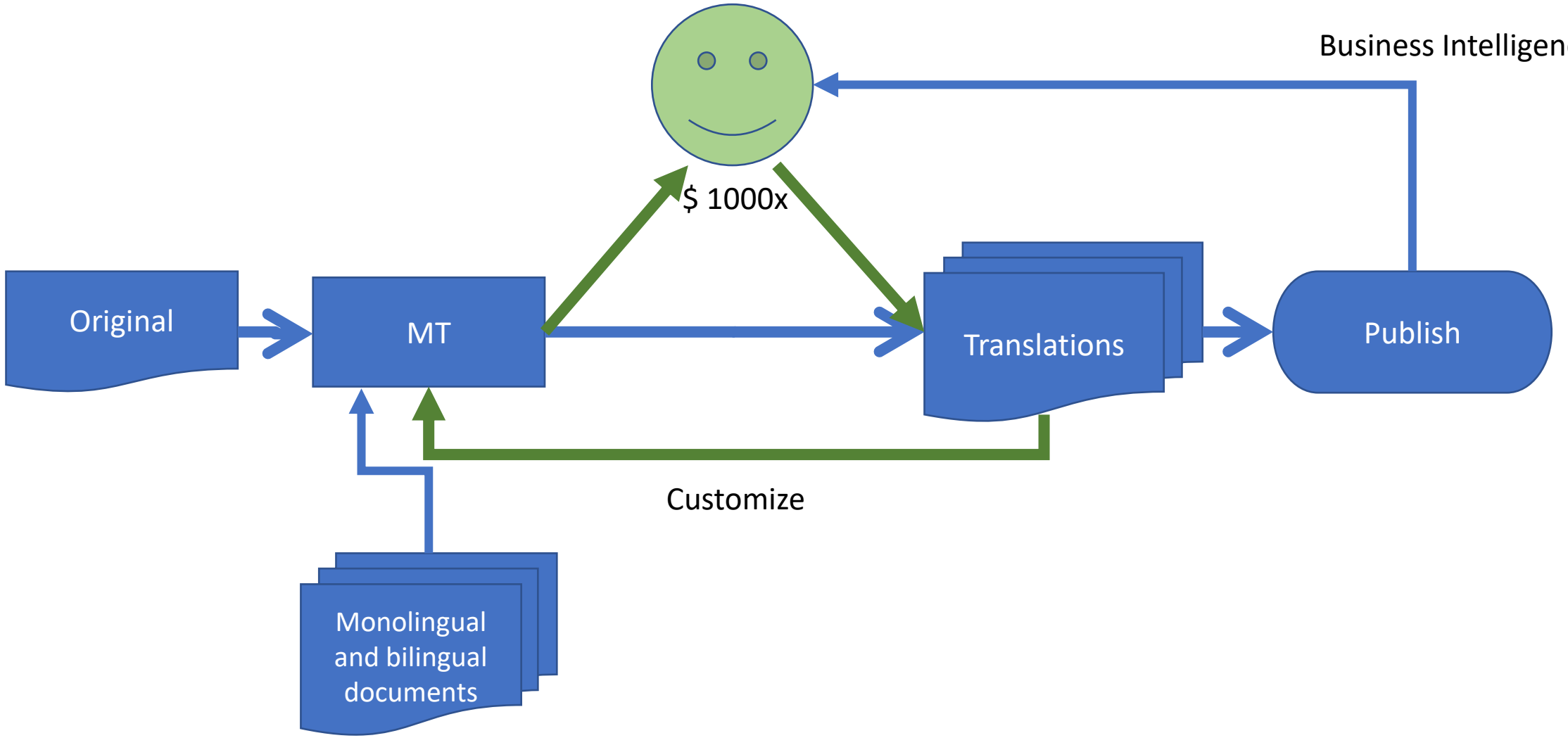
$280 is the runtime cost of customized Microsoft Translator.

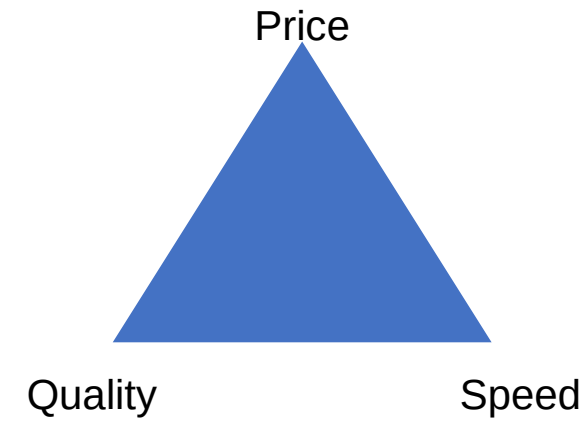
Uncustomized costs $70.

Post-Edit vs. Raw



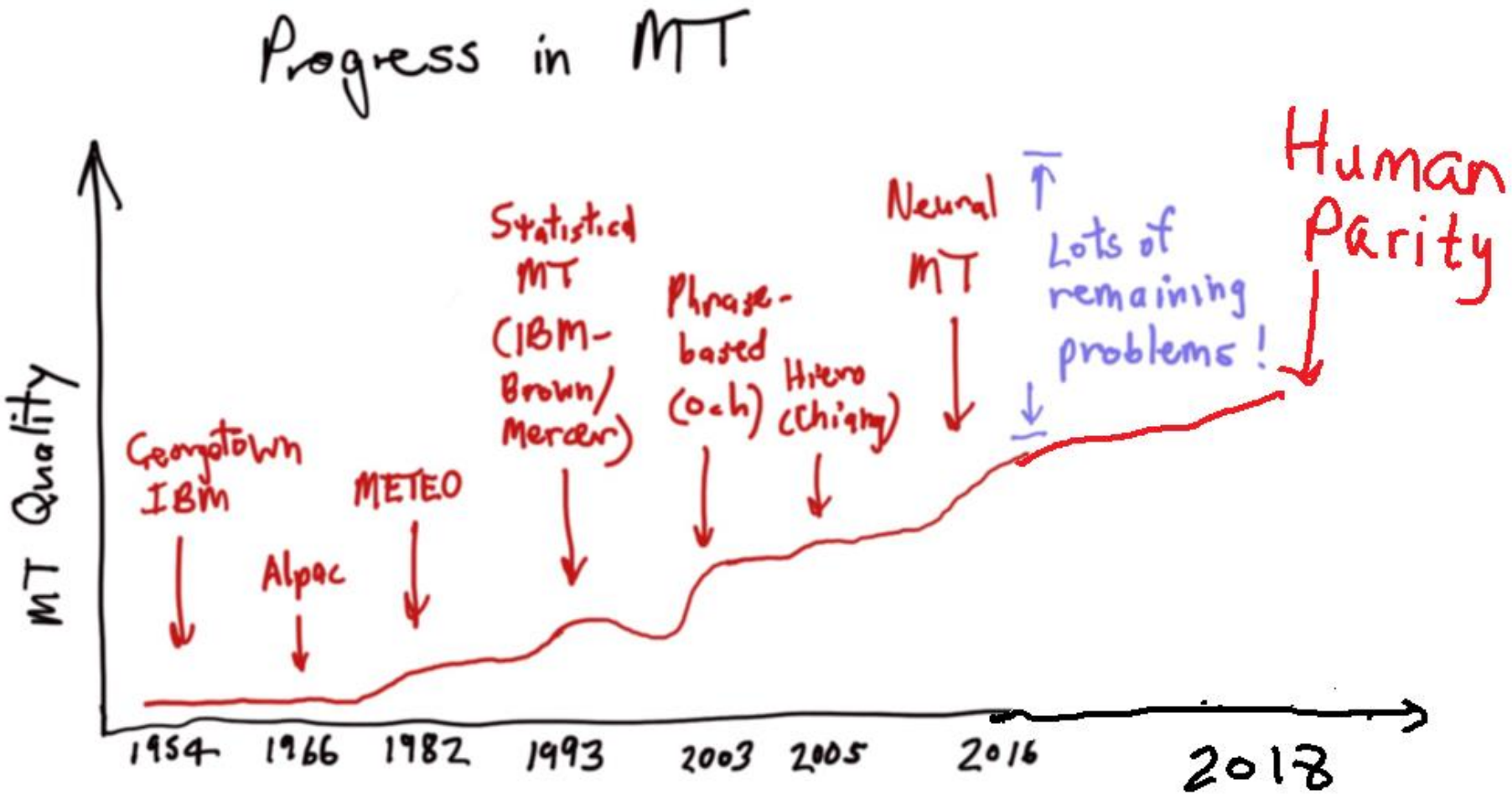
Post-Edit vs. Raw





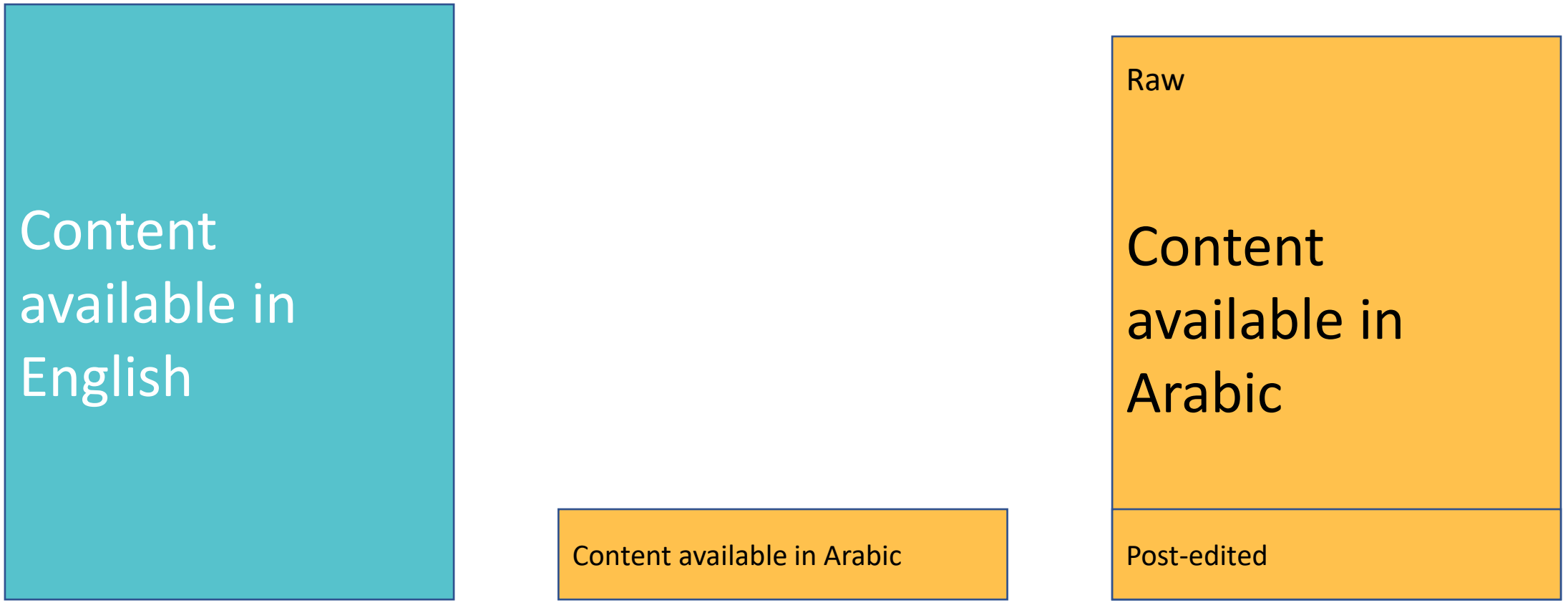
1000

1000 x more
1/1000 of the cost
not 1000 times worse



Edited from Chris Manning 2016

Extent of localization



Headquarter language

Today: 10%

Tomorrow: 95%

Where are we?

- Getting closer to parity
- Year over year
- For the most translated languages

What is missing to true human parity?

- Document context ← *not too far away*
- Non-zero chance of catastrophic mistakes
- Real-world knowledge

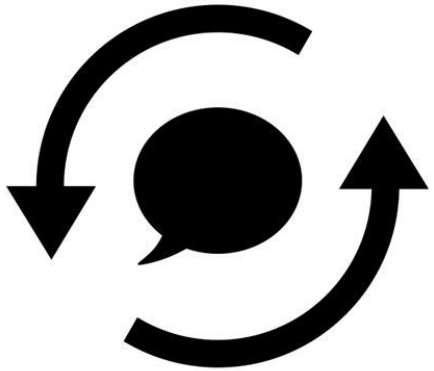
Whole Document Context

Today: Customization

- Train a custom translation system for the terminology and style of the enterprise or topic
- Good improvements
  (~ 10 point BLEU)
- Dictionaries
- Use previously translated prose for machine learning
- In domain, representing your style and terminology
- 20,000 sentences is a good start

Next

- Whole document context
- Cross-sentence transfer
- Sensory context



Example English - German

The sun was about to set. It was still shining brightly.

Die Sonne stand kurz vor dem Untergehen. Es leuchtete noch hell.

The “sun” in German has female gender. Second sentence uses neutral gender.

The sun was about to set, it was still shining brightly.

Die Sonne stand kurz vor dem Untergehen, sie schien noch hell.

Correct female gender for “sun” in the translation.

Translated by Microsoft

Catastrophic mistakes

Dutch original	Nederland en Australië stellen Rusland aansprakelijk voor zijn aandeel in het neerhalen van vlucht MH17.
Our original neural translation to English	The Netherlands and Australia are responsible for its share of the MH17 of the flight.
Today's neural translation to English	The Netherlands and Australia make Russia liable for its share in the taking down of flight MH17.

Translated by Microsoft

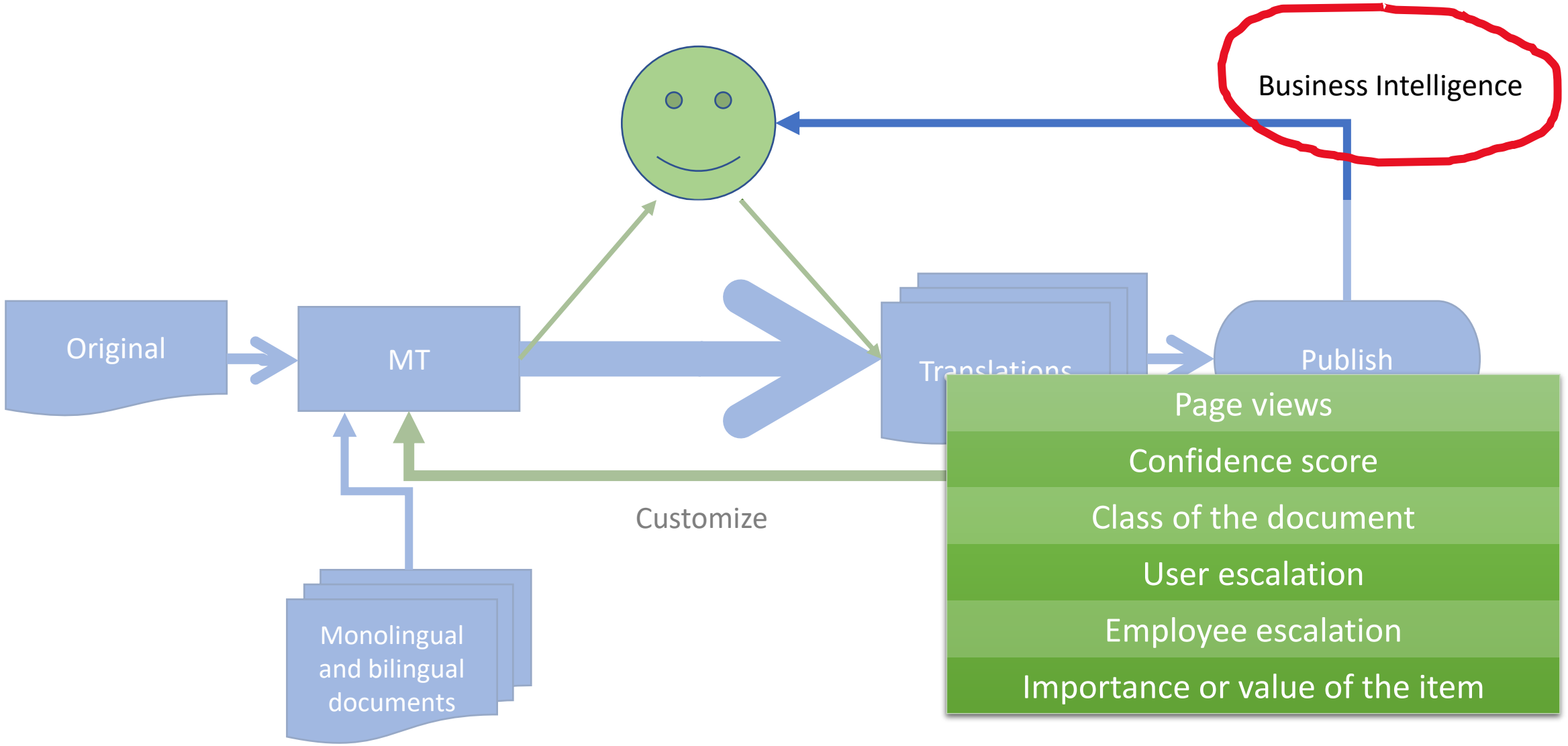
Catastrophic mistakes

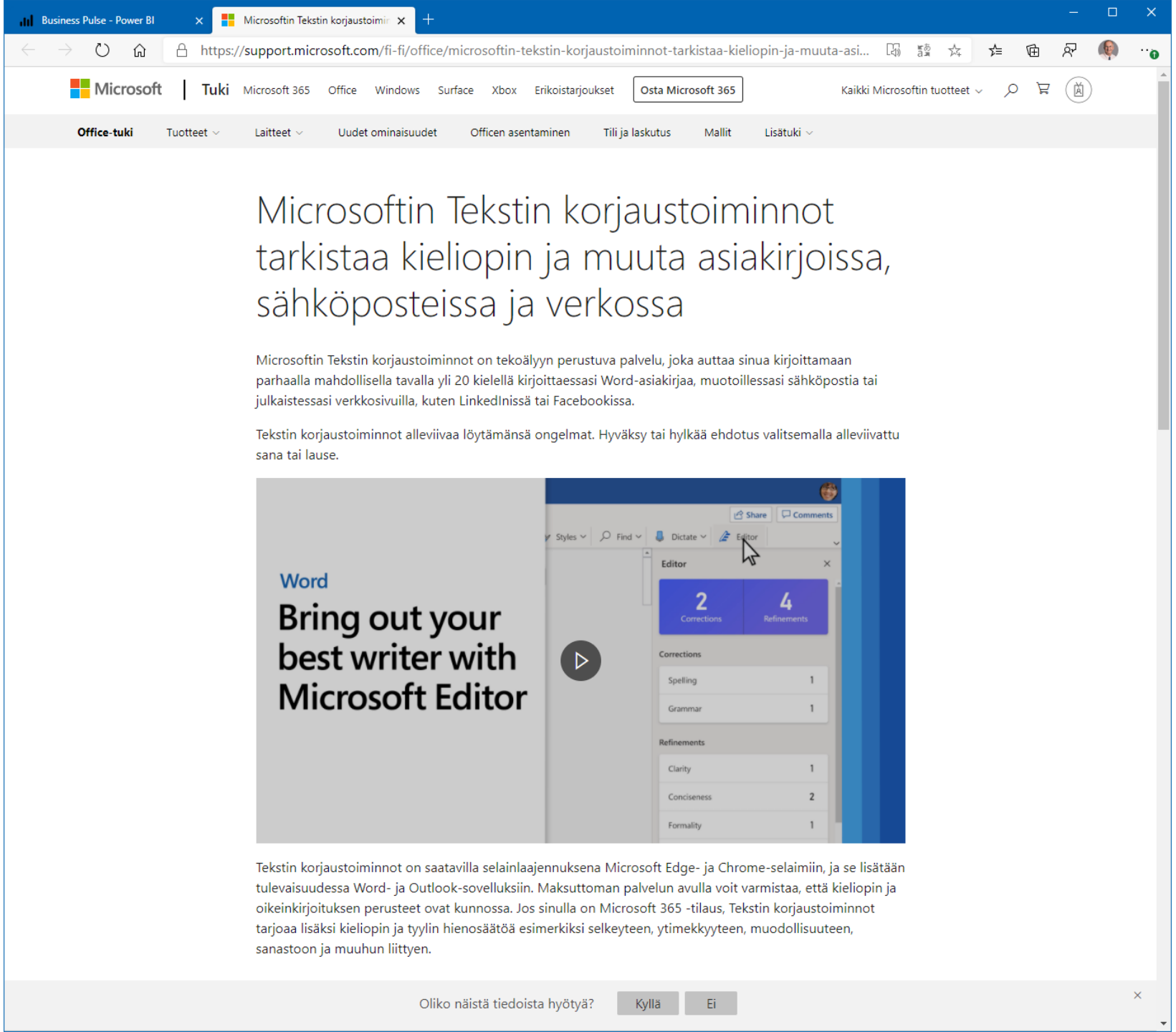
Inverted negation	The original says to not do something, the translation says to do something.
Changed number or unit of measurement	It may happen that a dimension indicator, currency or unit of measurement (inch, centimeter) gets incorrectly translated
Falsified person names and titles	It may happen that the wrong title is applied to a person, for example “King” where it should have been “Crown Prince”.
Religious figures or symbols placed in unfavorable context	Ambiguity in the training material can lead to a translation that shows a religiously significant figure or a relevant symbol in an unfavorable light.
Omissions and hallucinations	The machine translation may omit a relevant concept or add a concept that was not present in the original.

Machine Translation

Pros	Cons
Machines translate 1000 times cheaper than humans.	Machines can introduce translation mistakes that need to be corrected by humans.
Machines translate infinitely faster than humans.	
Machines translate with a quality that is approaching human translation quality.	

Post-Edit vs. Raw





ILTA-SANOMAT

found grammar errors
in a support article about
Microsoft Editor, a grammar checker.

Can you spot the grammar error?

Microsoft Finland:
We will keep monitoring the situation, as I think this encourages people to read the MSFT pages more carefully because of this error, and try to spot similar errors elsewhere.

Language sequencing

The class of document determines which languages raw MT is suitable for

Document class suitable for machine translation			
Rank	Class A	Class B	Class C
1. English – Spanish	✓	✓	✓
2. English – French	✓	✓	✓
3. English – German	human	✓	✓
4. English – Japanese	human	✓	✓
5. English – Korean	human	human	✓

Actual ranking FROM English

1	Italian
2	Chinese (Traditional)
3	Chinese (Simplified)
4	French
5	Japanese
6	Spanish
7	Danish
8	Hindi
9	Portuguese (Brazil)
10	Swedish
11	Arabic
12	German
13	Norwegian
14	Welsh
15	Dutch
16	Korean
17	Czech
18	Telugu
19	Turkish
20	Malaysian
21	Indonesian
22	Bulgarian
23	Croatian
24	Icelandic
25	Russian
26	Romanian
27	Greek
28	Hebrew
29	Thai
30	Tamil
31	Hungarian
32	Swahili
33	Vietnamese
34	Latvian
35	Maltese
36	Finnish
37	Ukrainian
38	Slovak
39	Filipino
40	Farsi
41	Slovenian
42	Urdu
43	Catalan
44	Polish
45	Serbian
46	Estonian
47	Marathi
48	Malagasy
49	Lithuanian

Ranked by delta to human translation quality

We compare the human evaluation of human translated sentences to the same source sentences machine translated and sort by the delta, smallest to largest.

Translated by Microsoft

Takeaways

- Increase the extent of localization
- Dare publishing raw MT - Think factor 1000
- Implement feedback and customization
- Make use of business intelligence
- Intra-company communication
- As a customer-facing product feature

- Interactive bots
- Analytics and data mining

The only job:

Remove the language barrier

chris.wendt@microsoft.com