BECOMING A MACHINE TRANSLATION COACH

LANGUAGE ASLING TECHNOLOGY

AI in the Language Industry

- Neural Machine translation (NN & RNN*)
- Gathering and cleaning of training data
- Quality assurance tools
- Estimating localization quality
- Helping to quickly select the best linguist for a particular job
- Forecasting & Planning
- Recommending engines for optimal workflow selection
- Voice recognition & Voice to text software

asian absolute

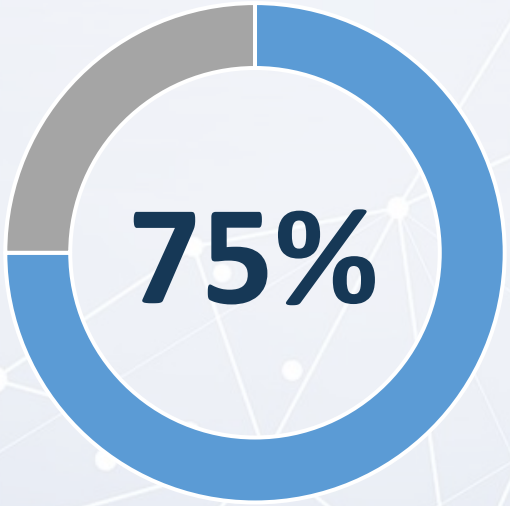
Machine Translation Benefits

- Lower price
- Human parity* can be achieved for certain dataset and language pairs
- Higher efficiency, increasing speed of translation
- MT integration into CAT tools
- Opening new markets / breaking down language barriers

asian absolute

2019

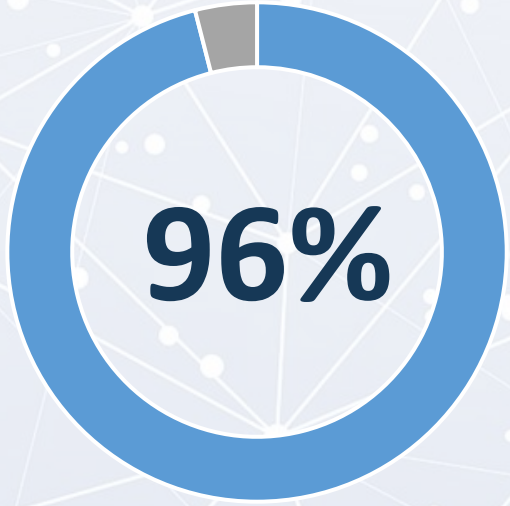
Approximately **14** languages are sufficient to reach about



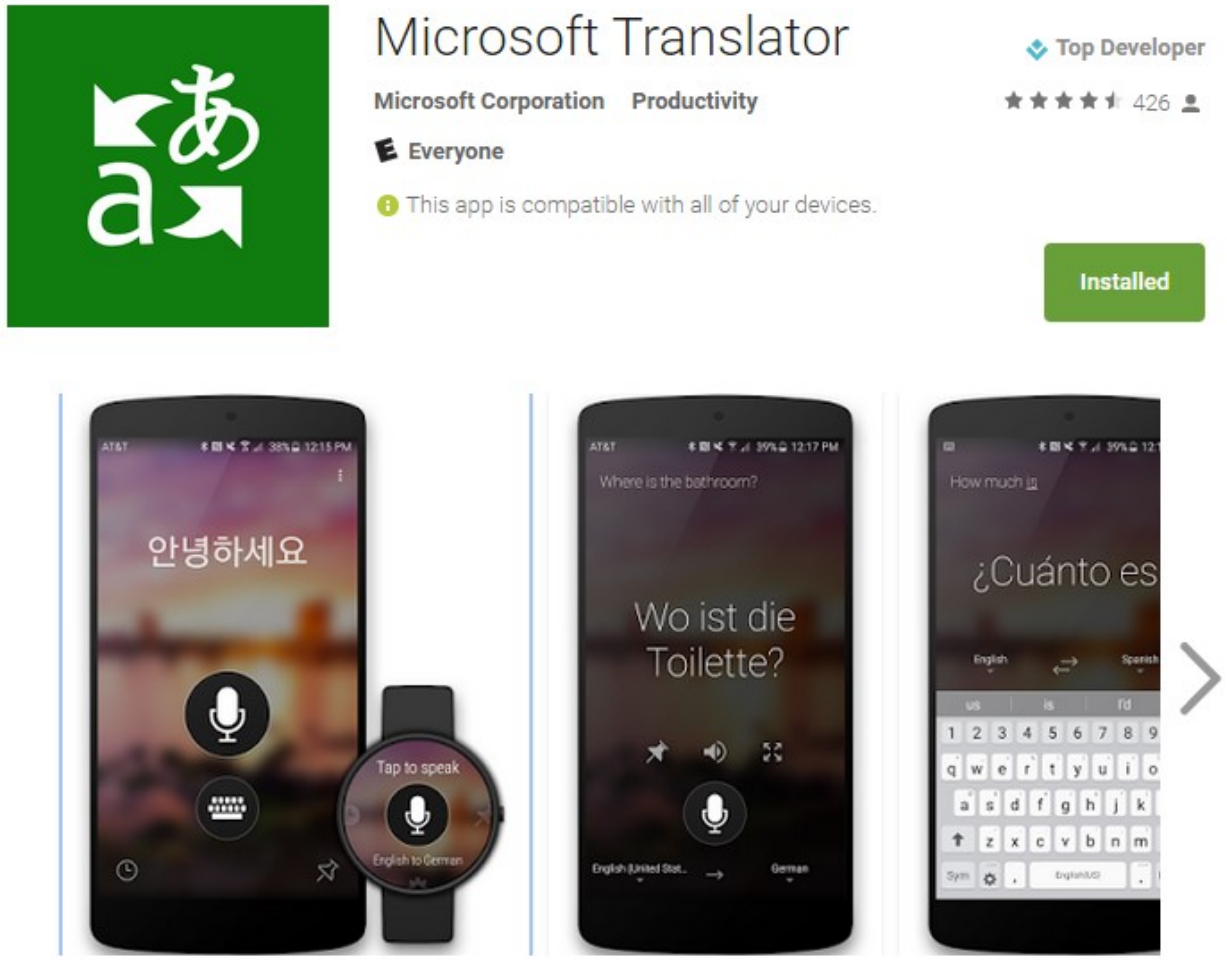
of the global Internet users

2027

Enterprises will need to translate into more that **60** languages in order to reach

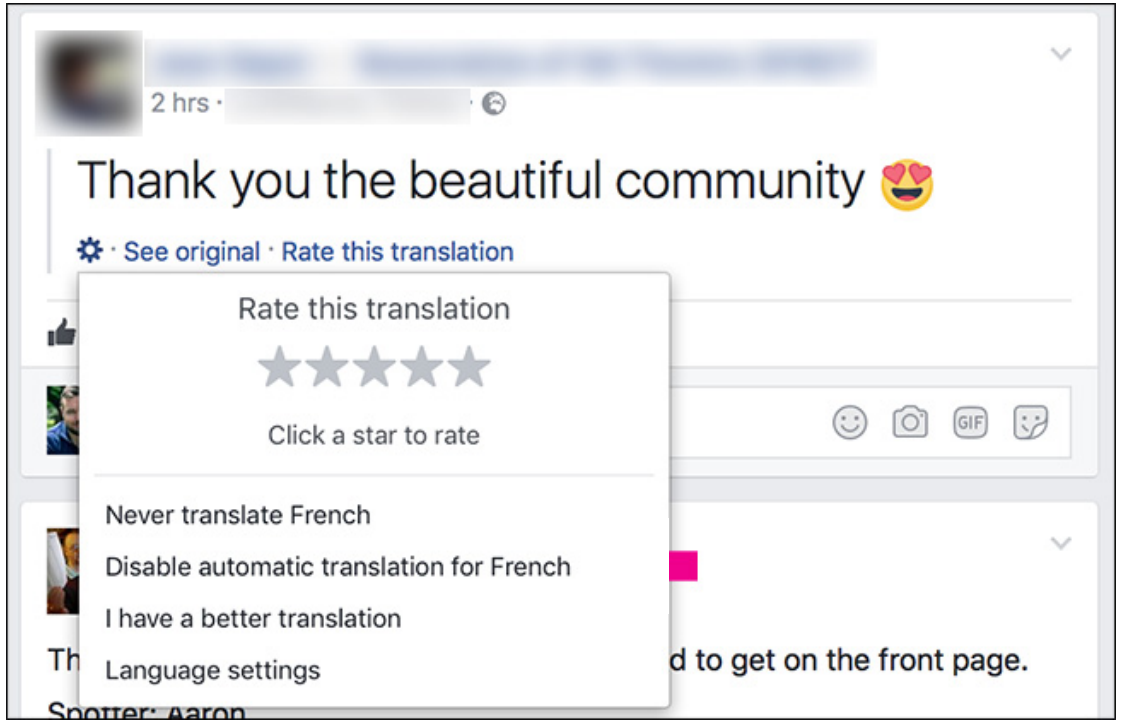


of the global Internet users



Microsoft has developed its own Translator app, which not only translates text, but also speech, images, and street signs. Microsoft's big breakthrough with this app is it can run offline, which has ideal real-world advantages for those traveling in areas with limited connectivity.

Facebook’s translation system underwent a major overhaul in the summer of 2017 and processes around 4-5 billion translations everyday using AI (LSTMN). AI takes into consideration the context, typos, abbreviation and the intent making for much more accurate results. The site’s “rate this translation” function also allows the neural network to update in real time through user input.



Machine Translation Challenges

There are still many challenges such as:

- Data scarcity and data quality
- Cross-contamination
- Information Security, Compliance Regulations, GDPR
- MT underperforming on certain language pairs and dataset
- Some more creative and abstract material

asian absolute

Machine Translation Services

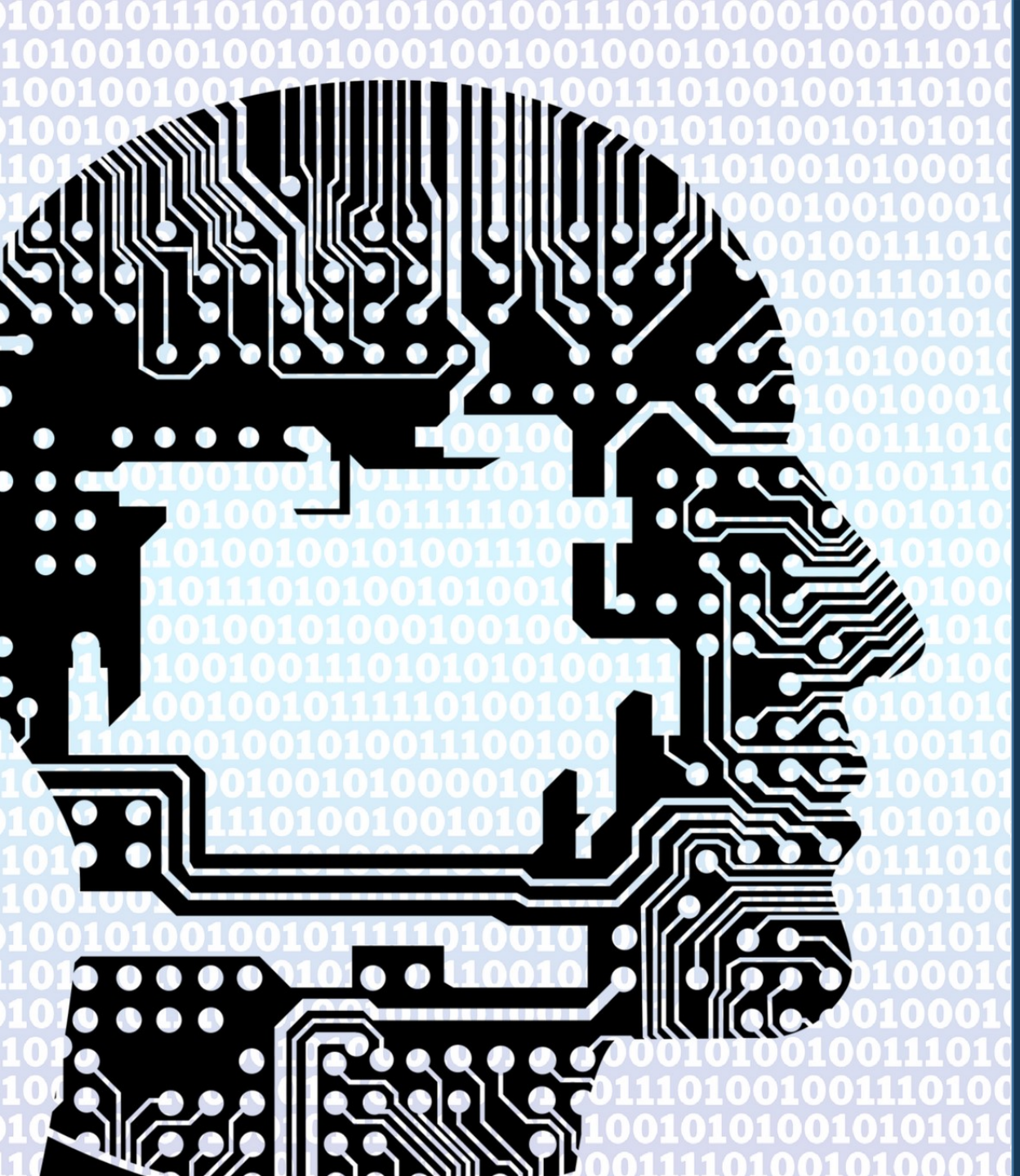
MT services we can easily offer without having the resources to invest or access to large amounts of data.

Quality Management:

- Cleaning the data
- Giving feedback to the engine
- Post editing

Human resources:

- Providing training to linguists to help transitioning their skills
- Providing data engineers training



Advanced data quality program is essential to using AI efficiently in translation.

The main factors to consider to assess the data quality to use for translation

Context
The type of data being cleansed and the purposes for which it is used

Storage
where the data resides

Data Flow
How the data enters and moves

Work Flow
How work activities interact with and use the data

Control
People responsible for managing the data

Continuous Monitoring
Processes for regularly validating the data

Data-cleaning Process

- **Manual** by a professional translator/linguist:

  For example, misalignment can be fixed by re-alignment or removing the non-matching data completely. For “bad” translations, re-translation is sometimes necessary.

- **Automatic** by running cleaning scripts and algorithms:

  For example, one phase is to check for long segments. By default, any segments with more than 40 words are rejected. This can be changed depending on the language combination and domain, but the default is 40 words.

asian absolute

CASE STUDY

Bilingual segments from the bilingual European Parliament Proceedings Parallel Corpus 1996-2011

1. We selected 2,000 EN-FR segments from the corpus

2. We trained a Statistical Machine Engine with this “raw” data unprepared and un-cleaned and then we ran a test text to be translated into French into this MT system.

3. We then scored the French translation using BLEU scoring system

asian absolute

CASE STUDY

BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate translation of text to one or more reference translations.

Example:

The BLEU metric ranges from 0 to 1. A perfect match results in a score of 1.0.

reference = [['this', 'is', 'a', 'test']

candidate = ['this', 'is‘, 'test']

score = (reference, candidate) 0.75

asian absolute

CASE STUDY

Bilingual segments from the bilingual European Parliament Proceedings Parallel Corpus 1996-2011

The score achieved

DataSet	Segments	BLEU
Europarl	2,000	0.48

CASE STUDY

Noises for “Data Processing Engineer” to fix
1. Wrong language (source)
2. Wrong language (target)
3. Untranslated (source)
4. Untranslated (target)
5. Short segments (max2)
6. Long segments (max40)
7. Name entities
8. Insertion error
9. Duplication error
10. Repetitions
11. Case sensitivity issue

Noises for “Linguists” to fix
1. Misaligned sentences
2. Misordered words (source)
3. Misordered words (target)
4. Have no corresponding translation in the corpus
5. Contain poor or indirect translations
6. Spelling errors
7. Name entities
8. Insertion error
9. Duplication error
10. Repetitions
11. Case sensitivity issue
12. Indirect translations or usage of metaphors
13. Translations that are not precise enough

CASE STUDY

Bilingual segments from the bilingual European Parliament Proceedings Parallel Corpus 1996-2011

We re-ran the same text into the Machine that was trained with the clean data and we re-evaluated with the same BLEU score system.

DataSet	Test Segments	BLEU
Europarl	2,000	0.73

AI tools augment rather than replace humans translators increasing translation suppliers’ ability to handle vastly increased volume while simultaneously meeting the stringent requirements of highly specialized translation in healthcare, law, engineering, and other technical verticals.

THANK YOU

Asian Absolute

Phone: +44 (0) 20 7456 1058
Email: info@asianabsolute.co.uk