VERINT

Transfer Learning for Machine Translation

Part 1 of N

Graeme West

PyData Edinburgh, 2019-11-07

Key takeaways

1. Choose this approach if you have specialised ‘jargon’, and ***lots*** of existing translations.
2. Training is easy, once you have a suitable pipeline
3. Data engineering is the key
4. Results? Wait for Part 2!

The problems at hand

1. Time
   - Want to release complete translations alongside each new version of our software
2. Money
   - We spend a lot of money on human translation
3. Scale
   - Market demanding new languages faster than we can deliver

The ‘solution’

1. G 文

2....

3.Profit??!

VERINT

Software strings ≠ prose

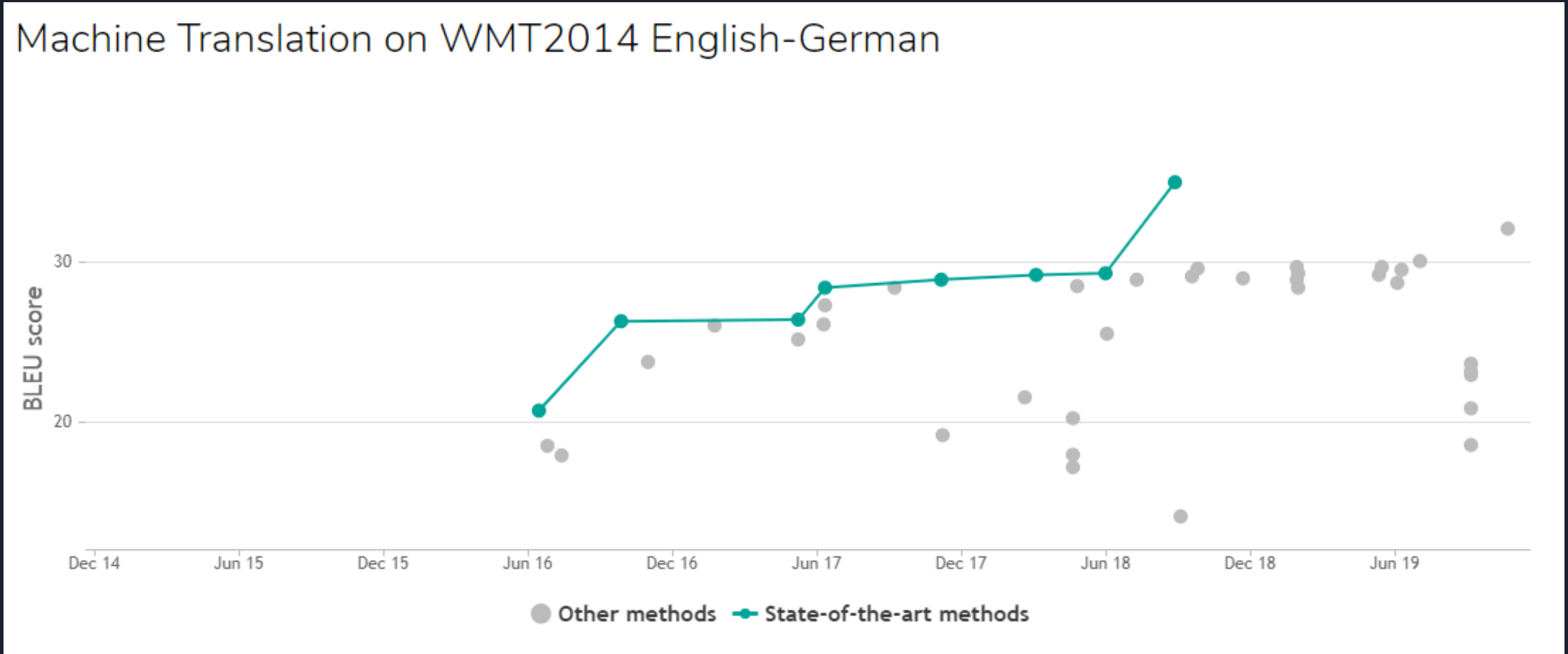
	English	German
Prose	Actors Orlando Bloom and Model Miranda Kerr want to go their separate ways .	Schauspieler Orlando Bloom und Model Miranda Kerr wollen künftig getrennte Wege gehen.
Software UI strings	An existing {0} for employee {1} has been removed. It starts from {2} to {3}.	Ein bestehendes {0} für Mitarbeiter {1} wurde entfernt. Beginnend mit {2} bis {3}.

Specialised terminology

Source EN	Reference Human Translation	General MT	Custom MT
“Time Off Bid”	Freizeitgebot	Auszeit	Freizeitgebot
“Adherence Mapping”	Planeinhaltungszuordnung	Festhalten Kartierung	Planeinhaltungszuordnung
“No assigned users”	Keine zugewiesenen Benutzer	Keine zugewiesenen Nutzer	Keine zugewiesenen Benutzer

VERINT

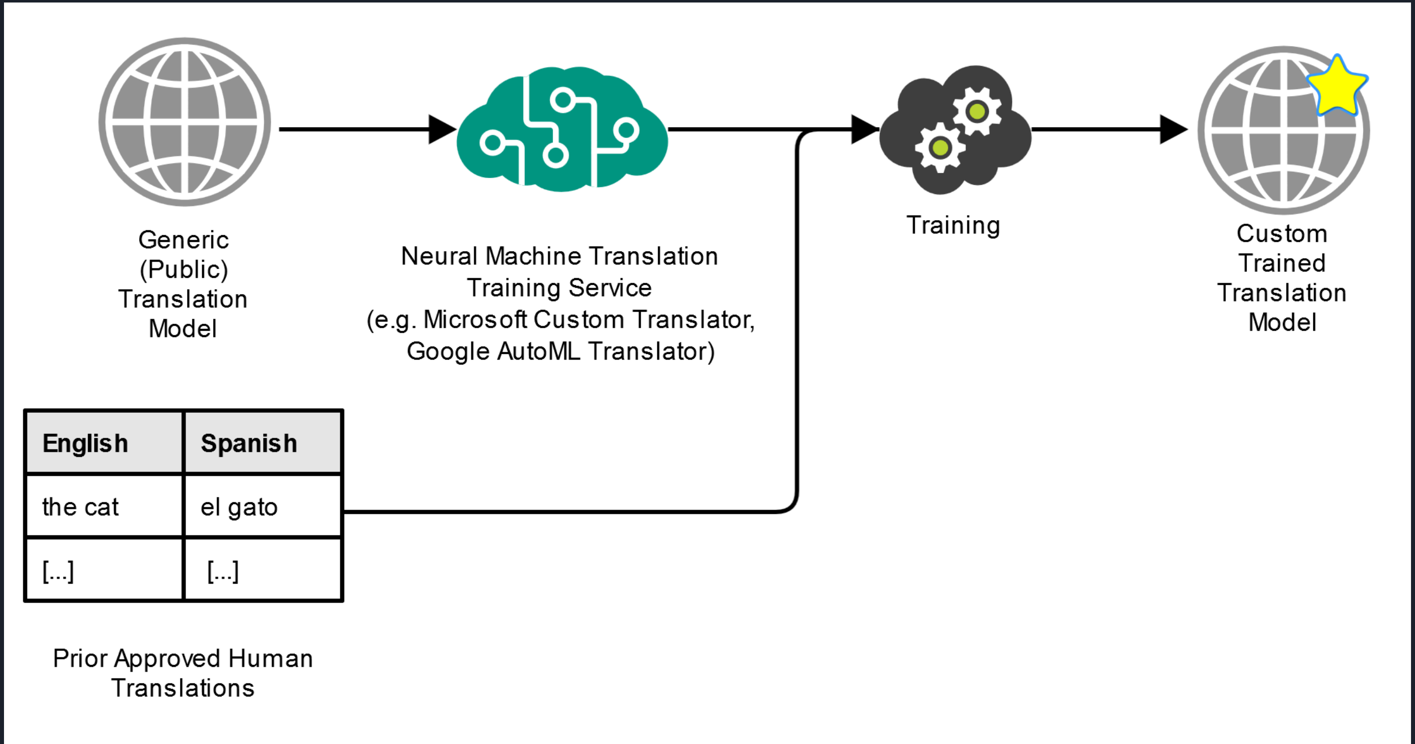
General MT not good enough(?)



Why transfer learning?

- We don’t have enough data/compute/PhDs/$ to build a large-scale model
- Let Google, Amazon, DeepL, & Microsoft spend the big bucks
- Virtuous cycle: we can review with humans and re-train later

The Process



Generic (Public) Translation Model
Neural Machine Translation Training Service (e.g. Microsoft Custom Translator, Google AutoML Translator)
Training
Custom Trained Translation Model

English	Spanish
the cat	el gato
[...]	[...]

Prior Approved Human Translations

Obligatory big numbers

- Training data:
  - English ⇒ German paired segments
  - ~2.2 million source words
  - ~200k segment pairs
- Train/test split:
  - ~99% train, 1% test (2500 segments)
- BLEU score improvement:
  - 39.65 (baseline)
  - 59.16 (custom MT)

Engineering

- Tools Used:
  - Google AutoML in experiments,
  - Azure Custom Translator in PoC
  - CrowdIn TMS
- ETL for training data = big challenge
- APIs have hard limits
- Need to work on data provenance and model management

VERINT

Results so far

- Able to provide instant ‘first pass’ translation for humans to approve later
- Costs lower than expected ($40/m chars)
- Notable improvement in sentence coherence and terminology preservation
- Downside: can’t ‘own’ our models (yet)
- Next steps: native speaker evaluation