VERINT®

Transfer Learning for Machine Translation

Part 2 of 2

Graeme West

PyData Edinburgh, 2020-01-09

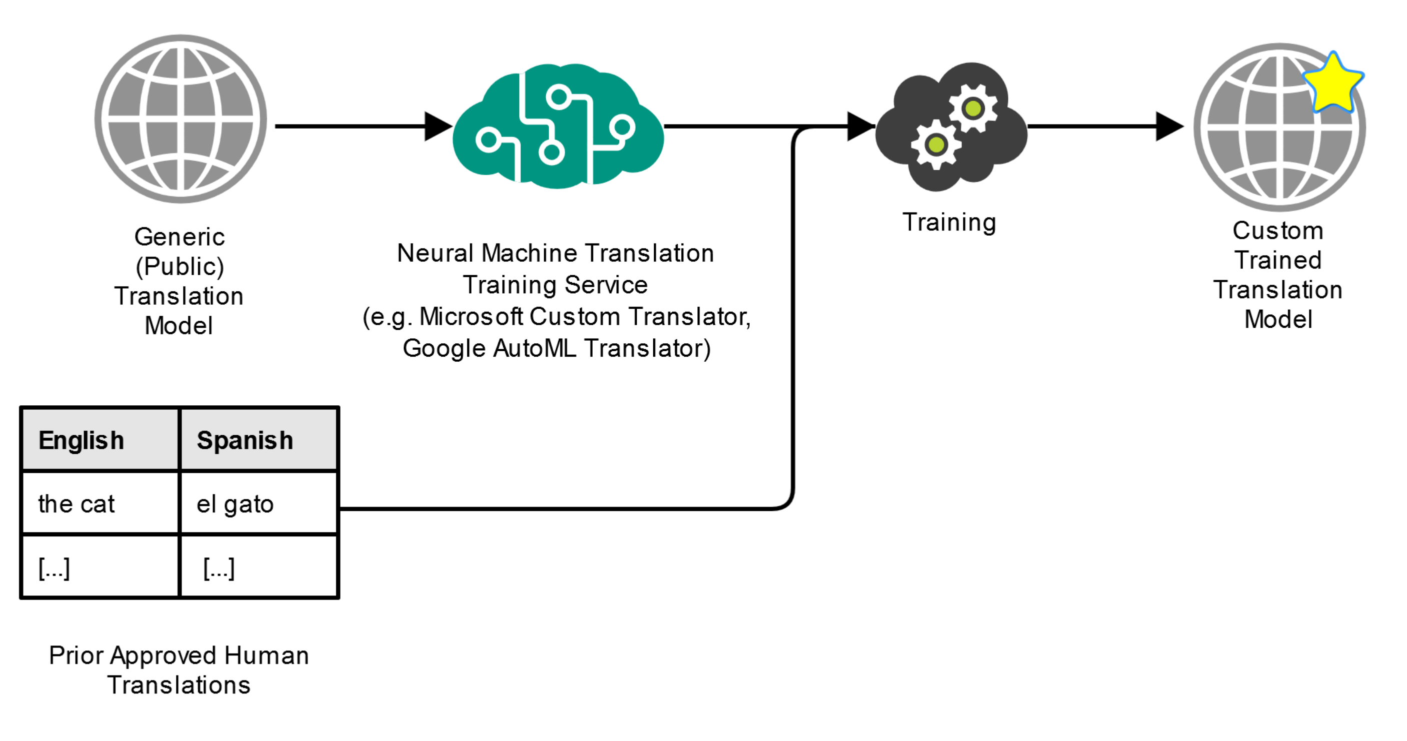
Summary of Part 1

1. We translate lots and lots of user interface text and online help for software products
2. Human translation is expensive & slow
3. We have custom product vocabulary, which general MT engines perform poorly on
4. We started a proof-of-concept to on *trained*, *custom* machine translation and transfer learning

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}.

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

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

Approach to evaluation

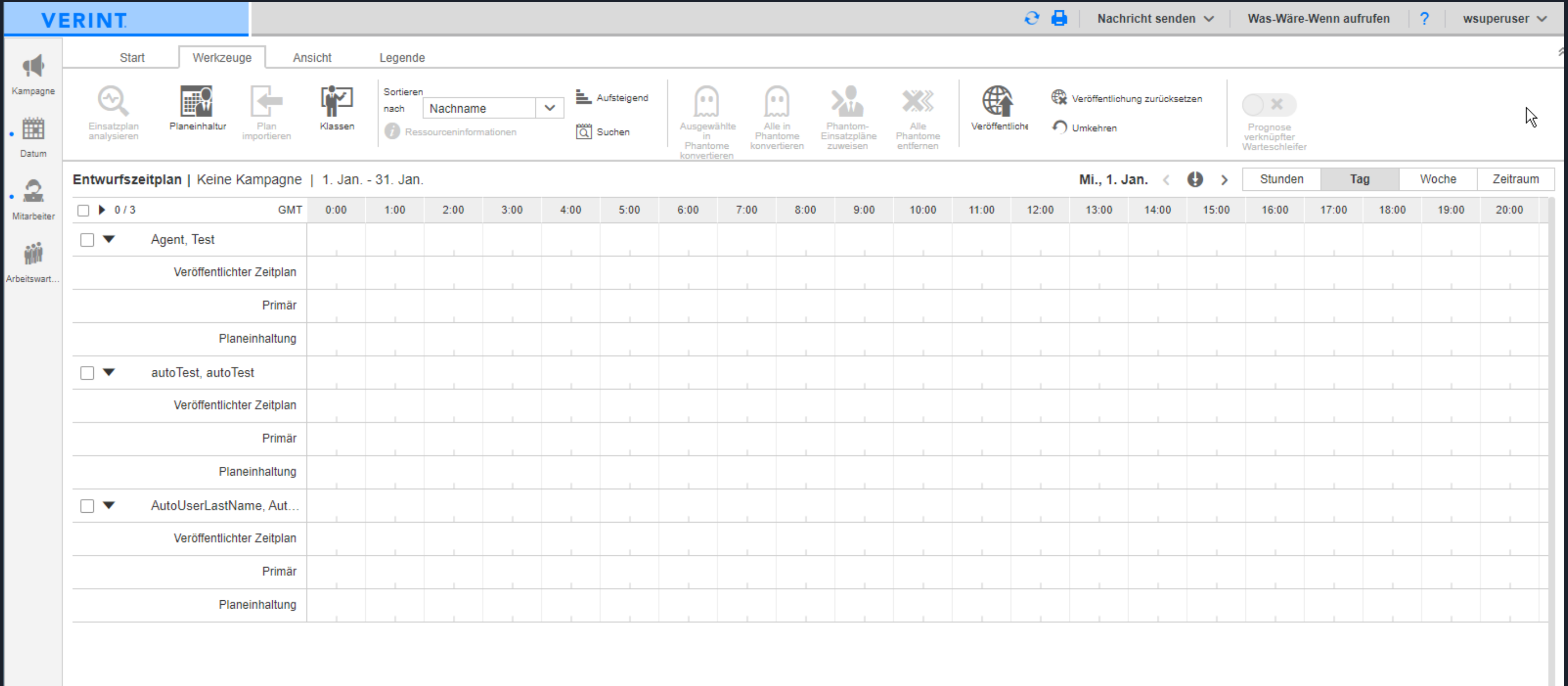
- Qualitative - Human native speakers
- Quantitative - Calculated BLEU scores

Human native speaker evaluation

Research question: how close are these to the translations you would pick yourself?

Participants must be domain experts *and* native speakers.

Parallel demo systems



Entwurfszeitplan | Keine Kampagne | 1. Jan. - 31. Jan.

□ Agent, Test
Veröffentlichter Zeitplan
Primär
Planeinhaltung

□ autoTest, autoTest
Veröffentlichter Zeitplan
Primär
Planeinhaltung

□ AutoUserLastName, Aut...
Veröffentlichter Zeitplan
Primär
Planeinhaltung

VERINT

Human evaluation results

- “I am impressed how good the quality is!”
- New vocabulary was less well translated (not enough training data)
- ‘Sense’ of words sometimes slightly off:
  - Is ‘View’ a noun or a verb?
  - ‘Queue’ translated as ‘physical queue’ (e.g. in a supermarket) rather than a conceptual waiting list

Quantitative evaluation

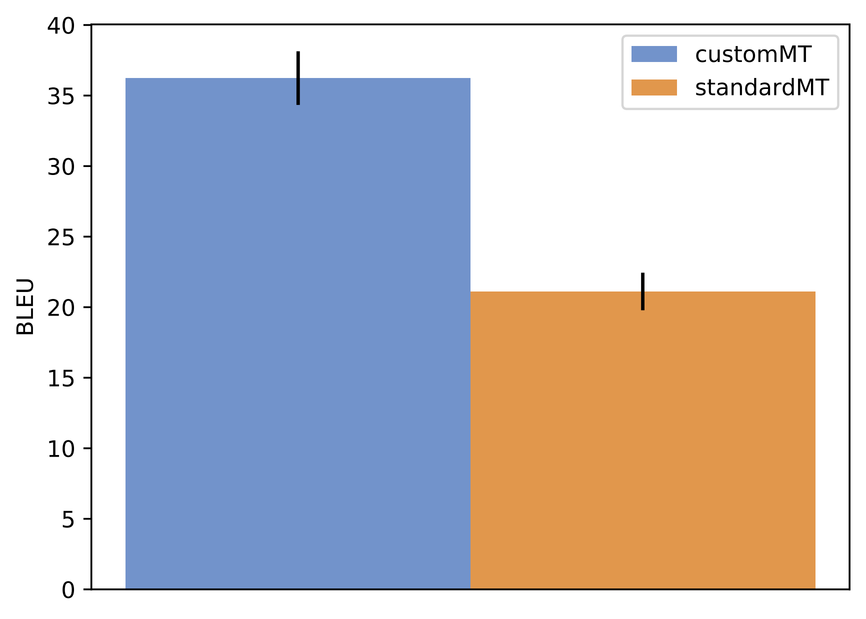
- Inputs:
  - Candidate custom MT translation
  - Generic MT translation
  - Reference human translation
- Tools
  - CompareMT Python package
- Caveats:
  - Any form of automated evaluation is flawed: no single ‘right’ translation exists.

Quantitative evaluation -BLEU

- BLEU (Bilingual Evaluation Understudy Score)
  - Compares 1-grams to 4-grams
  - Focuses on precision, rather than recall *“How many N-grams in the reference translation showed up in the candidate translation?”*
  - Theoretical max score: 100
  - State of the art MT score: 30-50

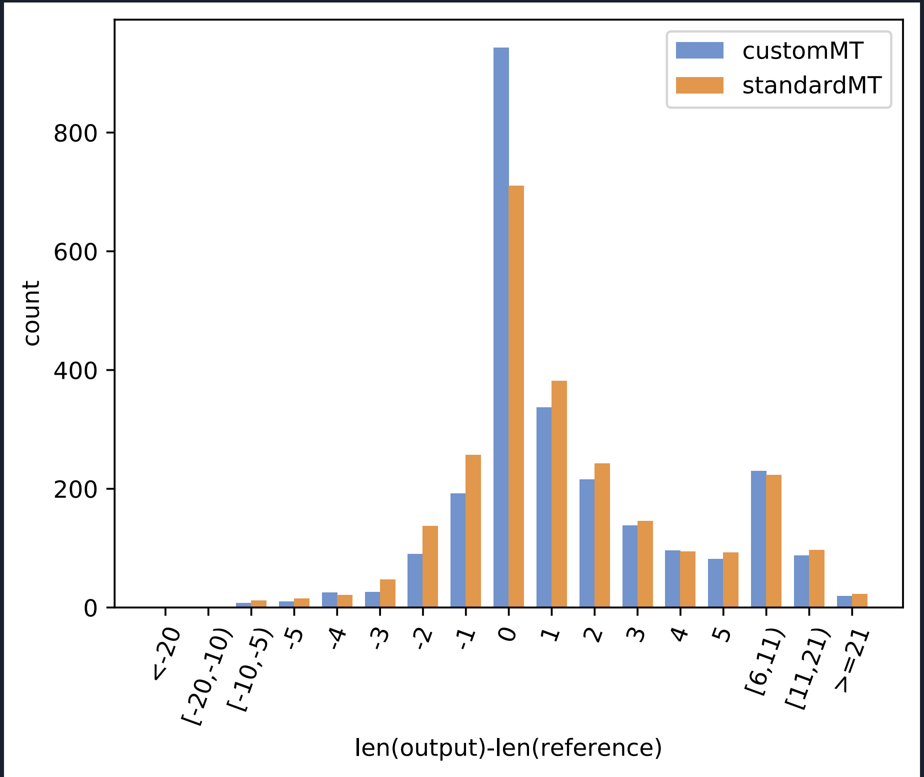
VERINT

Overall BLEU score improvement



	customMT	standardMT	Win?
BLEU	36.2320	21.0915	s1>s2
	[34.3368,38.1381]	[19.7839,22.4446]	p=0.0000

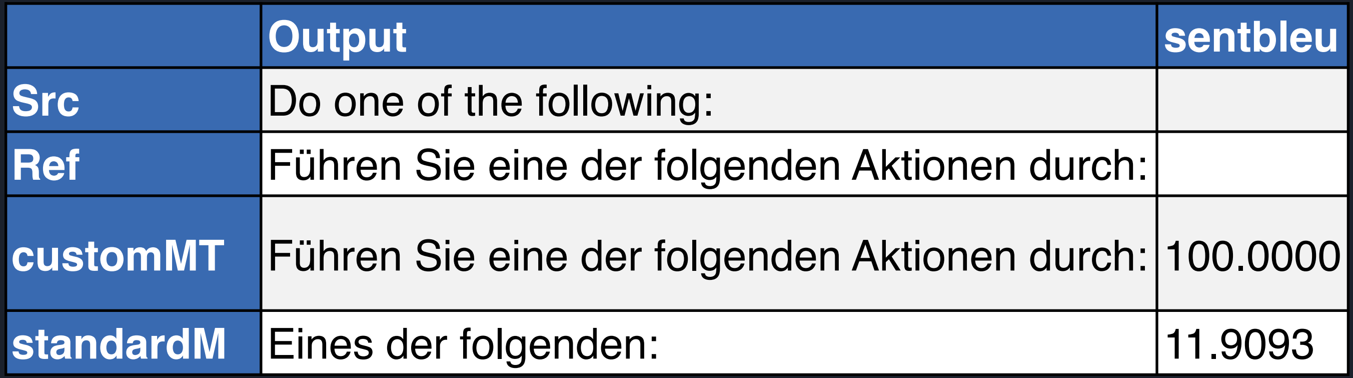
Sentence length concordance



Custom MT shows closer concordance in sentence lengths to human reference translation

Issues

- Separating training and test data with NLP is hard
  - Phrase recursion problem - overfitting?



	Output	sentbleu
Src	Do one of the following:	
Ref	Führen Sie eine der folgenden Aktionen durch:	
customMT	Führen Sie eine der folgenden Aktionen durch:	100.0000
standardM	Eines der folgenden:	11.9093

- New terminology lacks adequate training data to ‘shift’ the output in a particular direction

Next steps

- Roll-out to additional languages and locales
- Can we provide ‘part of speech’ annotations to improve MT results?
- Refinement of the re-training process
- Solution to the ‘recursion’ problem