

Opportunities for Machine Learning in the Translation Process

Aleš Tamchyna, Dalibor Frivaldský, David Čaněk
September 6, 2018



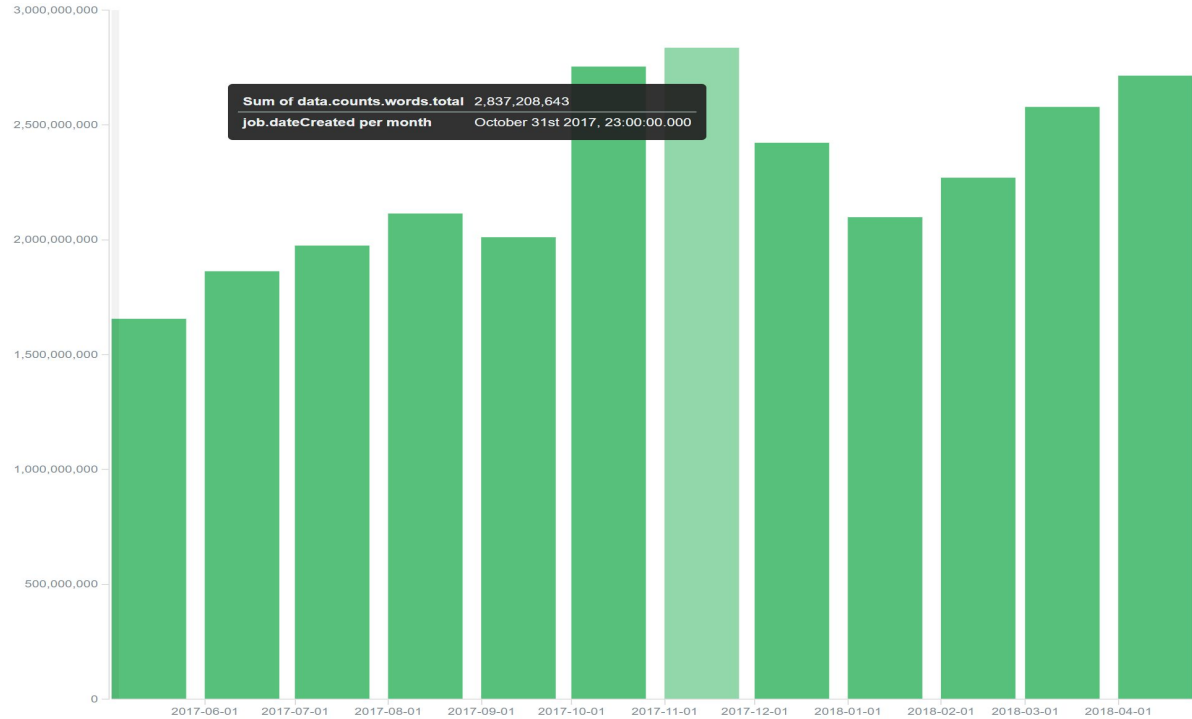
Outline

- **About Memsorce**
- Demo
- Identification of Non-Translatables
- MT Quality Estimation
- Towards Translation Automation
- Conclusion

About Memsource

- Cloud-based translation management system
- Translation editors (browser and desktop)
- Global customer base
- Founded in 2010
- Currently about 90 people (around 30 developers)

Words processed monthly



Some Interesting Facts

We support 300 locales, including languages like Inuktitut, Nuosu, Oriya, Iu Mien, Sami, etc.

Sami is spoken by 3 thousand people.



We have a customer that wanted us to support **Klingon**, so we do.



The Government of Nunavut has become a Memsource customer and asked us for some improvements for the **Inuktitut** language.



Outline

- About Memsource
- **Demo**
- Identification of Non-Translatables
- MT Quality Estimation
- Towards Translation Automation
- Conclusion

Demo

Outline

- About Memsource
- Demo
- **Identification of Non-Translatables**
- MT Quality Estimation
- Towards Translation Automation
- Conclusion

Why start with AI now?

- New possibilities opened thanks to the advances in deep learning
 - We are still discovering what can be done with these new tools
 - Space for new, innovative features
- Practical perspective:
 - Tools are stabilizing
 - Best practices emerge
 - Community, openness

Machine Learning in Translation

- Look at the whole process, identify tasks where AI can help
- First (“simple”) feature: detection of **non-translatable segments**
- Machine translation **quality estimation**
- Project management automation based on **content profiling**
- Overall goal: move **towards automated translation**

Identification of non-translatable segments

Web Editor Edit Tools Format Document Help

B I U X₂ X² ← → ✓ Split Join 🔍 → <> ¶ A|B

Filter Source Text Filter Target Text Aa Clear

#	Source: en	Target: cs			
1	16	16	✓	100	⋮
2	16 mm	16 mm	✓	100	⋮
3	john.doe@gmail.com	john.doe@gmail.com	✗	99	⋮
4	Madrid	Madrid	✗	99	⋮
5	Go to Madrid.		✗		⋮
6	http://www.memsource.com/	http://www.memsource.com/	✗	99	⋮
7	new_user_id	new_user_id	✗	99	⋮
8	<xref:System.Linq.ParallelEnumerable.WithMergeOptions%2A>	<xref:System.Linq.ParallelEnumerable.WithMergeOptions%2A>	✗	99	⋮
9	Spálená 51, 11000 Praha	Spálená 51, 11000 Praha	✗	99	⋮

Non-translatable segments

- translators have to attend to non-translatables...
 - ...and simply copy them to the target
- tedious and repetitive
- heuristics
 - no letters, only numbers/punctuation
 - does not cover enough
- can we identify them with ML?

An easy task?

1

2.1

`{{my.LPFormSubHead}}`

Outlook 2007

Shift+Enter

An easy task?

1



2.1

depends, maybe not for English-Czech: 2,1

{{my.LPFormSubHead}}



Outlook 2007



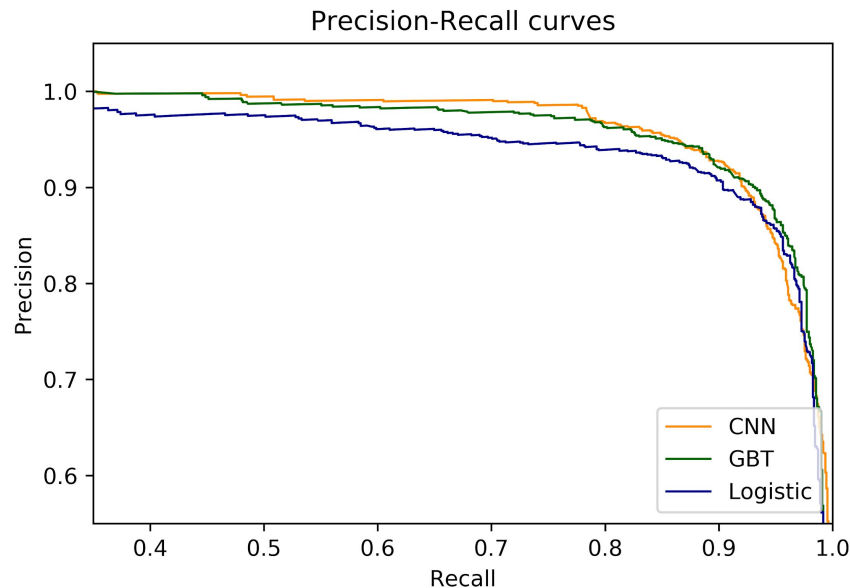
Shift+Enter

UMSCHALT+EINGABETASTE

AI-based NT detection

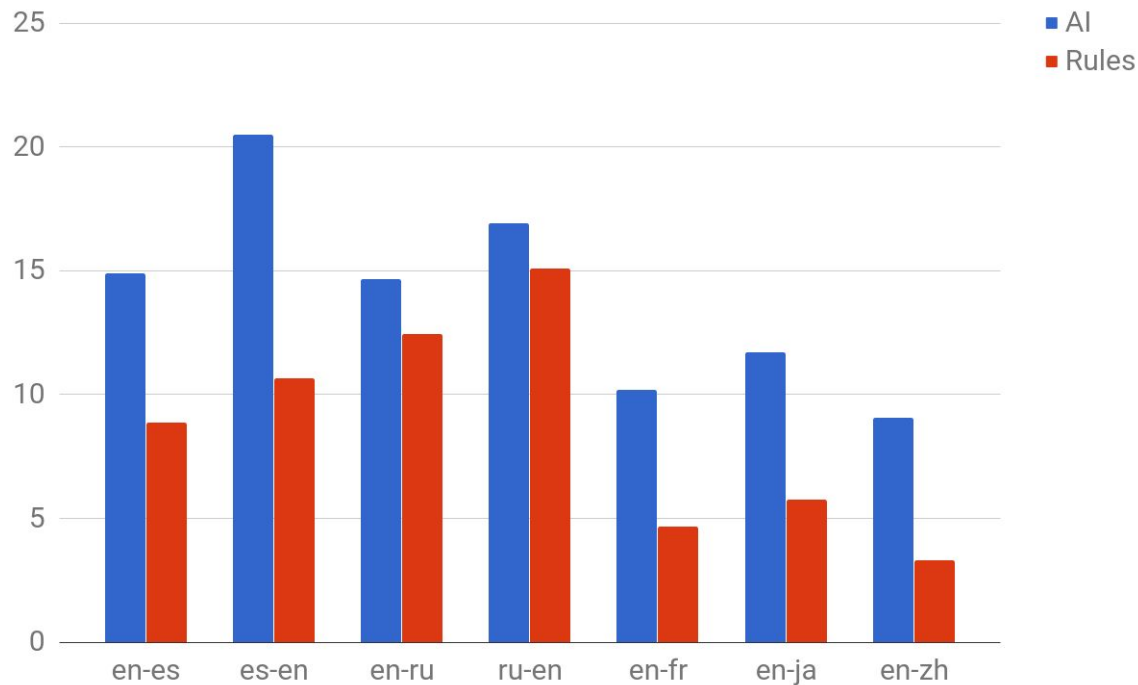
- binary classification problem
- train on historical data
 - which segments were changed, which stayed the same?
- be conservative
 - tune a threshold for classifying a segment as NT (high precision ~98%)
- each language pair is different
 - we support over 200 language pairs
 - most of them have a separate model

Evaluation of models for NT detection



- **character-level CNN**, gradient-boosted trees (xgboost), logistic regression
- very close in performance
- best model for us: CNN, precision stays ~1.0 the longest

Effect of AI on NT coverage



NTs: User Feedback

- generally positive:
 - savings
 - allows to offer more competitive prices
- complaints:
 - inconsistency
 - false positives
- Based on production data, AI-based NT detection is several times more accurate than the rule-based solution.

Outline

- About Memsource
- Demo
- Identification of Non-Translatables
- **MT Quality Estimation**
- Towards Translation Automation
- Conclusion

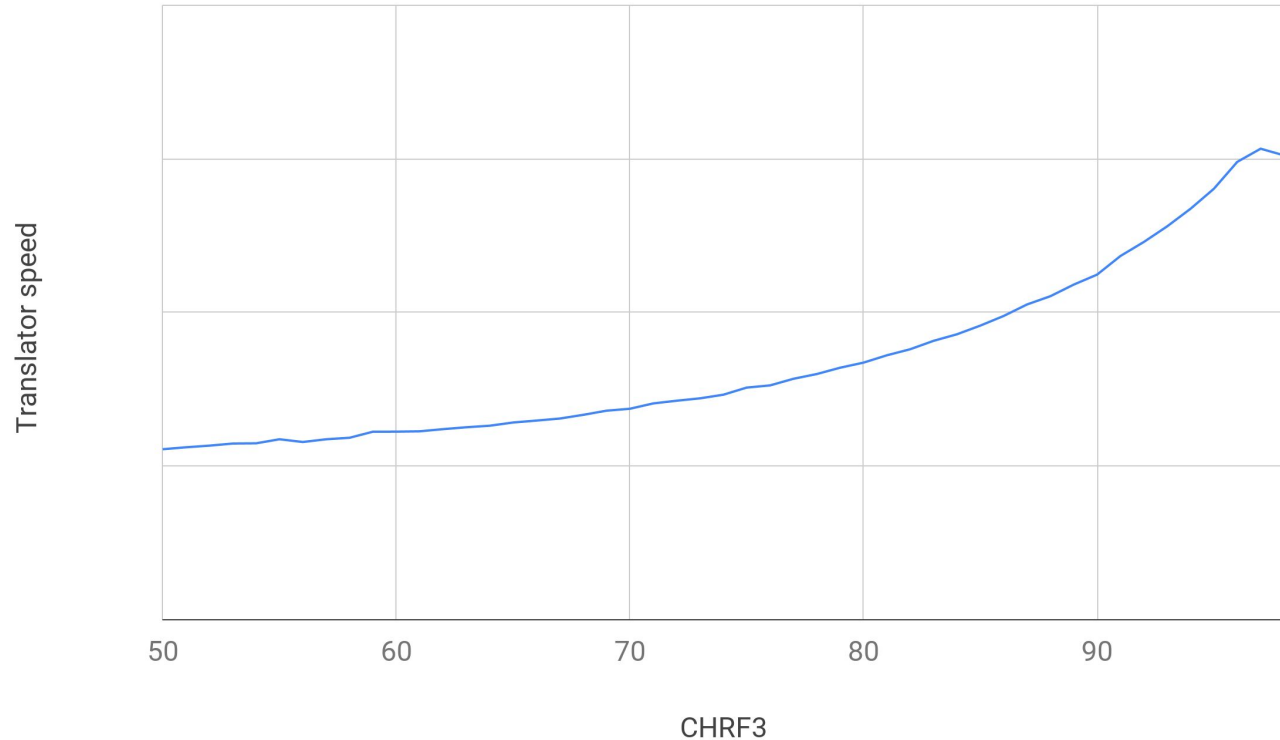
Is MT replacing human translation?

- Exciting result: human parity in Chinese-English news translation
 - more on that tomorrow, don't miss the keynote!
- Translators are still essential.
From an MT researcher's perspective, MT still struggles with:
 - coherence
 - terminology
 - style
 - ...
- But crucially: translators make sure that the correct **meaning** is conveyed
 - ambiguity, wording, emphasis, clarity, intent,...

Machine Translation in CAT

- The most “obvious” machine-learning based feature
 - human translation reduced to **post-editing** MT outputs
- Neural MT can produce high-quality translations
 - conventional wisdom: NMT outputs are typically **fluent and grammatical**
...but they can contain serious **errors which are easy to miss**
- Does it really save time?
 - ...that depends largely on MT **quality**
 - outputs which require no changes are still quite rare

MT Quality and Translation Speed



MT Post-Editing: The Problem

Translators have to go through lots of low-quality translations only to find the small percentage of cases where MT can actually help them.

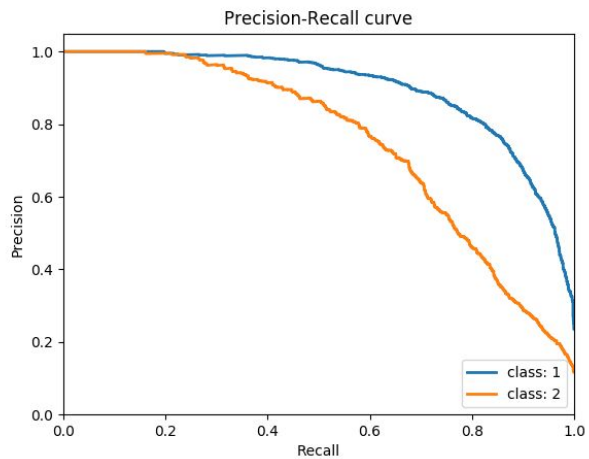
MT Quality Estimation

- a well-studied problem (see WMT QE shared tasks)
- general setting:
given a source sentence and MT output, judge the quality of the translation
- our approach:
 - sentence-level classification task
 - categories (version 1):
 - perfect MT
 - near-perfect MT
 - everything else

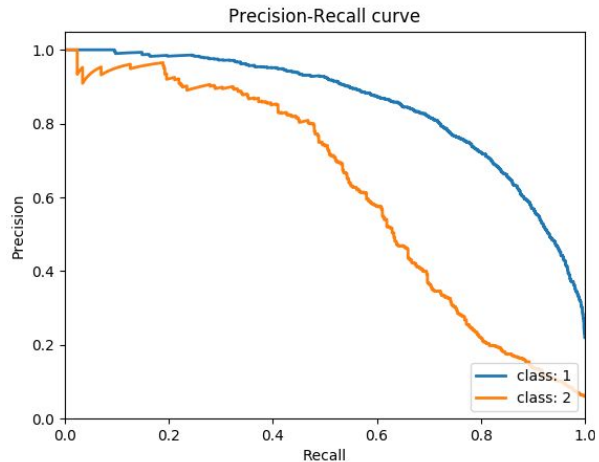
MT QE: Use cases

- **predict translation cost ahead of time**
- **score as a hint for translators**
- only show MT of sufficient quality to translators
- recognize suspicious human translations

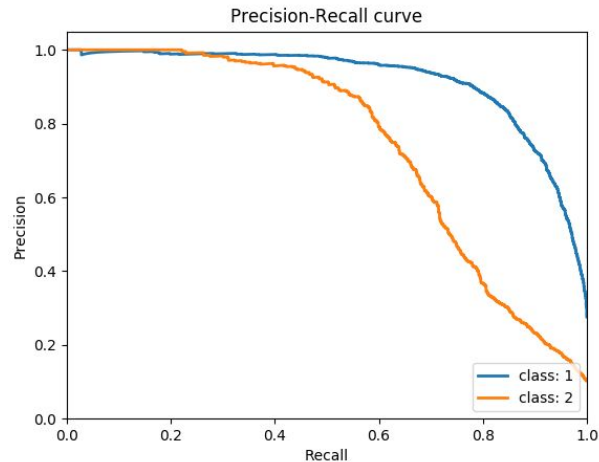
MT QE: Evaluation



Czech-English



English-Czech



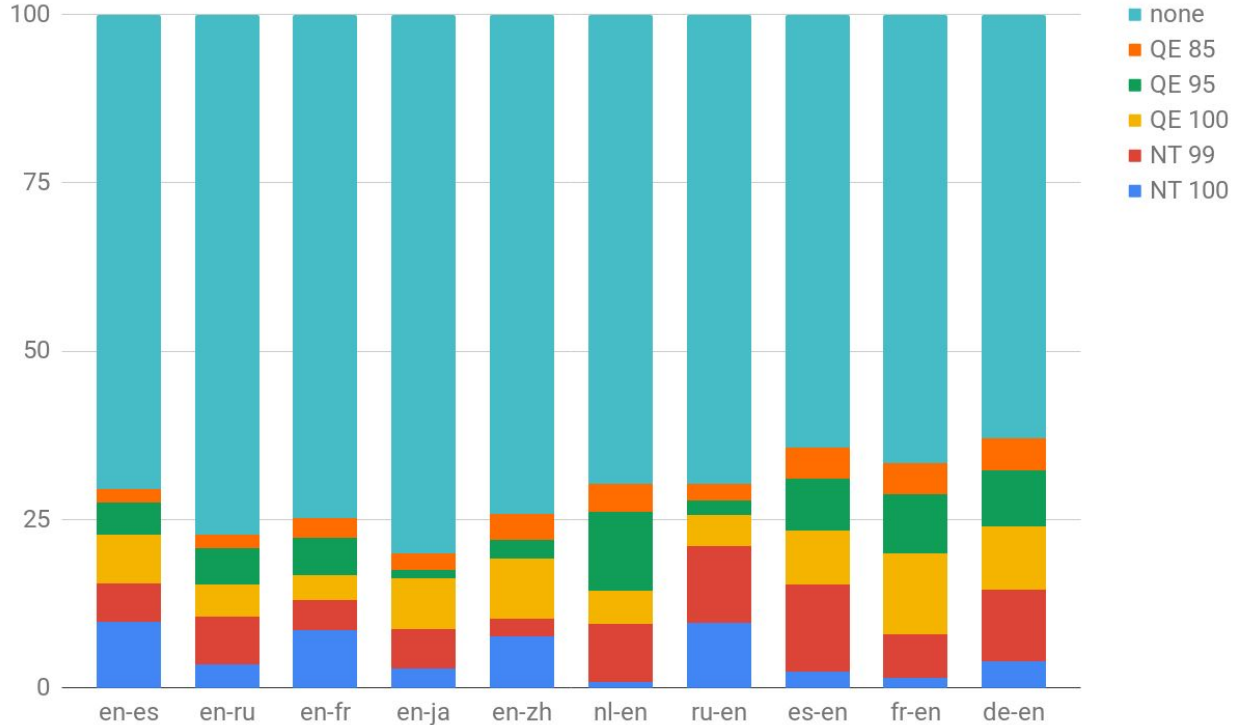
Russian-English

class 1: perfect MT
class 2: near-perfect MT

MT QE: Integration in Memsource Cloud

- 72 language pairs supported
- MT engine agnostic
- status: currently in private beta, to be released soon

NT Detection and MT QE: Segment-level Results



Outline

- About Memsource
- Demo
- Identification of Non-Translatables
- MT Quality Estimation
- **Towards Translation Automation**
- Conclusion

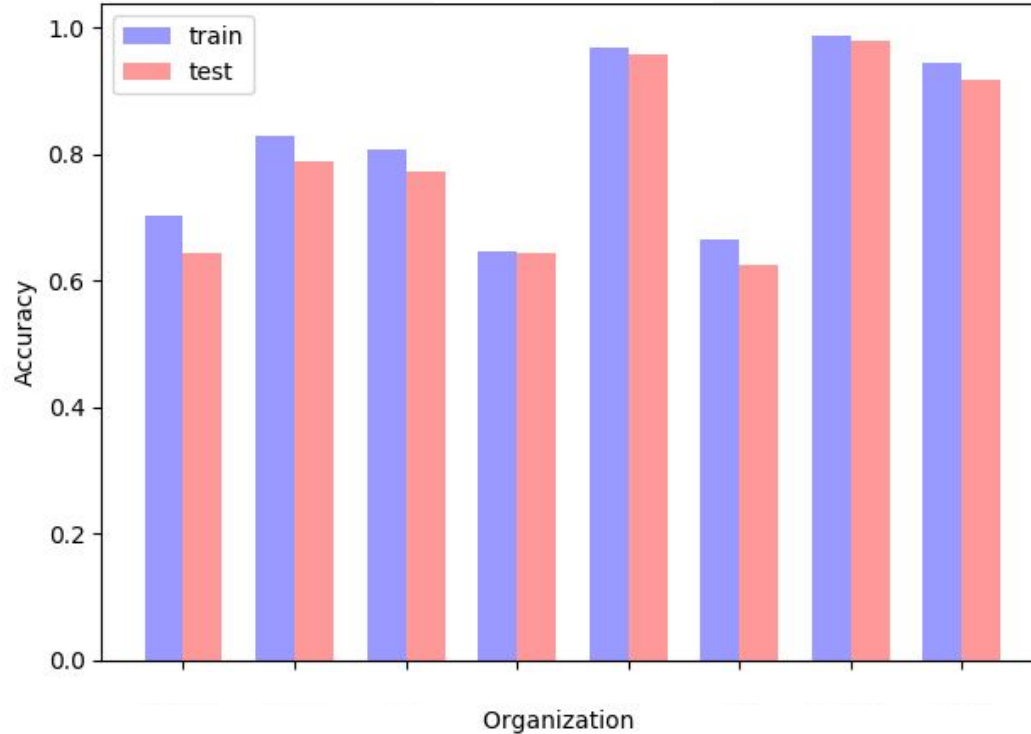
Towards Automated Translation

- Project management is time-consuming
- Repetitive tasks:
 - assign suitable **translators** based on domain expertise, experience,...
 - select appropriate **translation memories**
 - attach **term bases**
 - ...
- If we understood the content better...
 - ...we could automate some of these.
 - Solving these tasks is only the **first step** towards automation.

Content Profiling: Initial Steps

- First feature: automatic **domain identification**
 - Organizations manually assign documents to domains
 - Automatically suggest domain assignments
 - Classify unassigned documents
- Going forward:
 - Use learned representations to find matching translation memories,...

Domain Identification: Results



Conclusion

- AI in translation is not just MT
- ML models applicable in novel, interesting, useful ways
 - Detection of non-translatables
 - MT QE
 - Content profiling
 - ...?
- Translation industry is moving towards ever better automation

We're Hiring!

- AI researchers
- Software developers
- ...and more!

Check out <https://www.memsource.com/careers/>

Thank You!

Questions?