Neural Machine Translation: achievements, challenges and the way forward

Barbara Rychalska

b.rychalska@mini.pw.edu.pl

24.11.2019

Politechnika Warszawska

FINDWISE
SEARCH DRIVEN SOLUTIONS

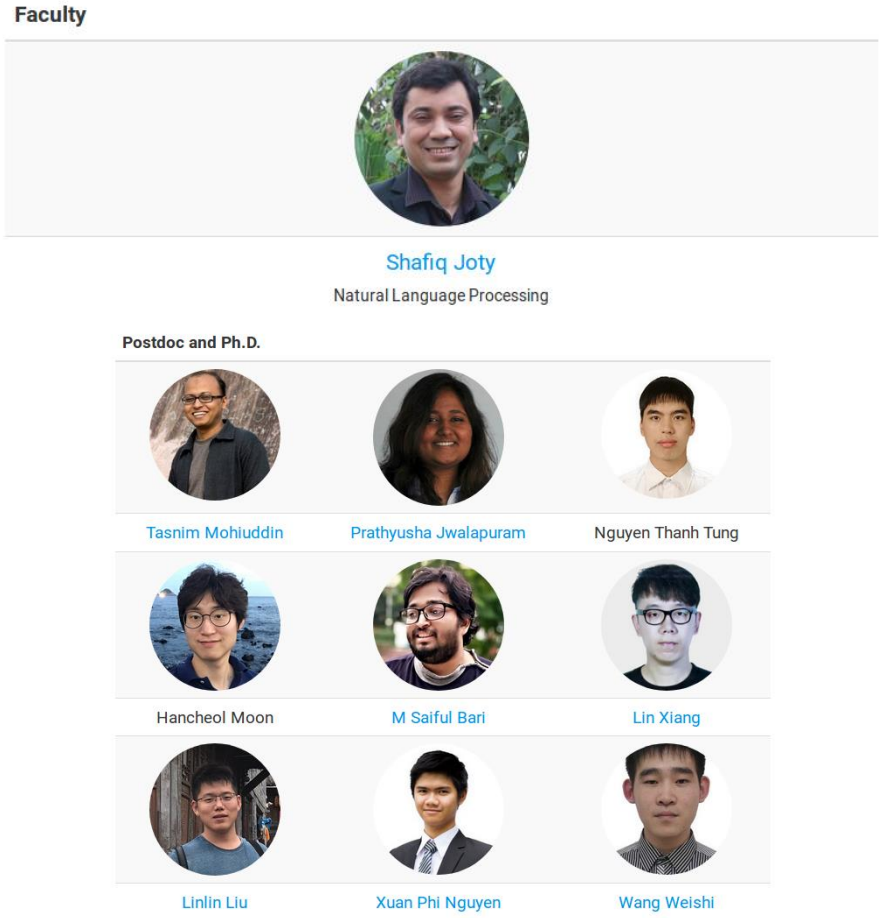
NANYANG TECHNOLOGICAL UNIVERSITY

MI DATA LAB

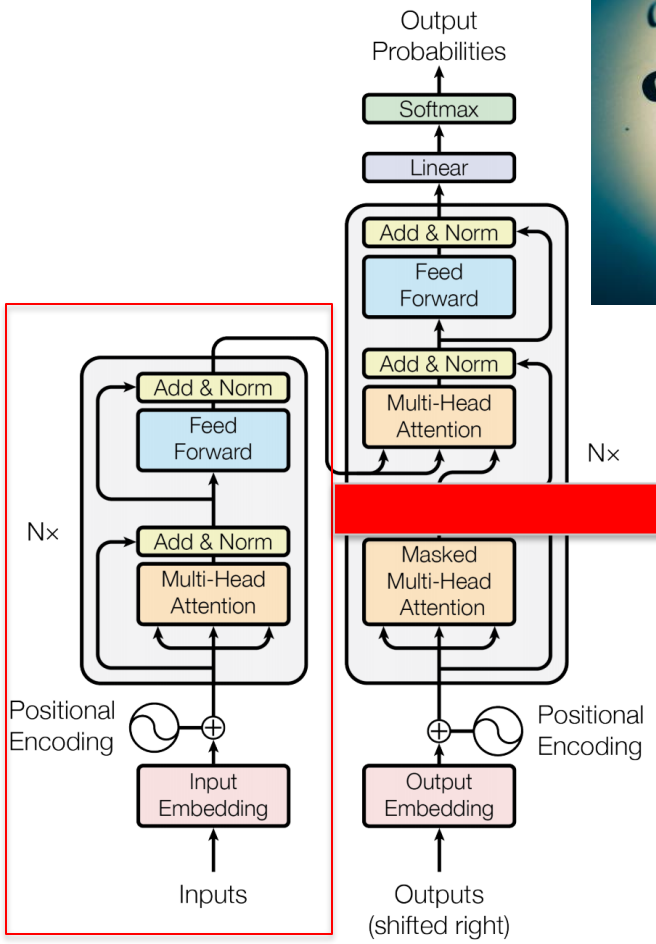
What this presentation is about

- Approach to Neural Machine Translation in cooperation between Warsaw University of Technology and Nanyang Technical University Singapore

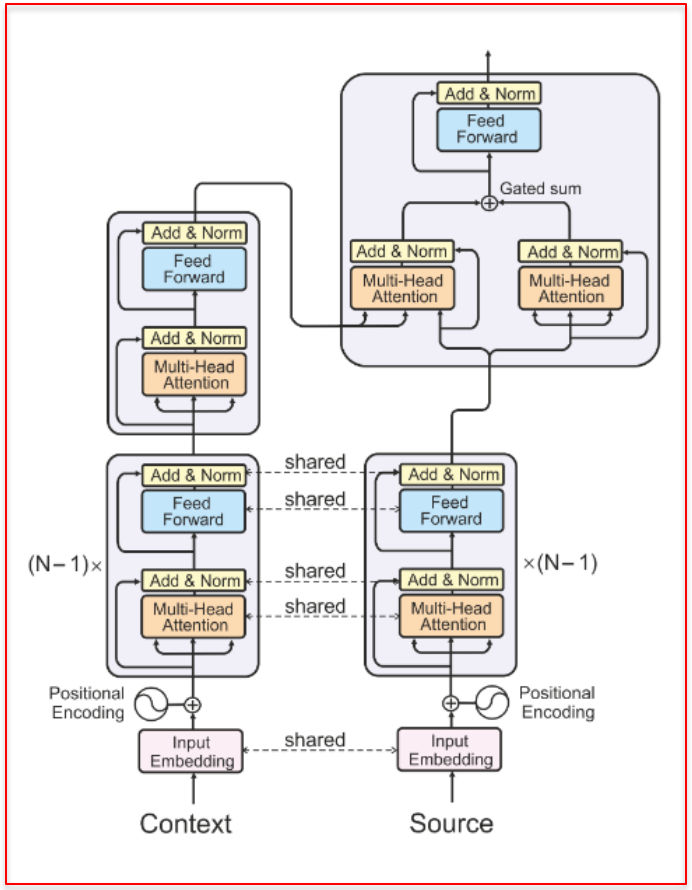
- Goal: answer questions
  - What we can do in NMT now
  - What we cannot do in NMT now
  - What can we do to make it better
  - What do we need to pay attention to



Sentence Level Translation vs Document Level Translation



Sentence-level Transformer

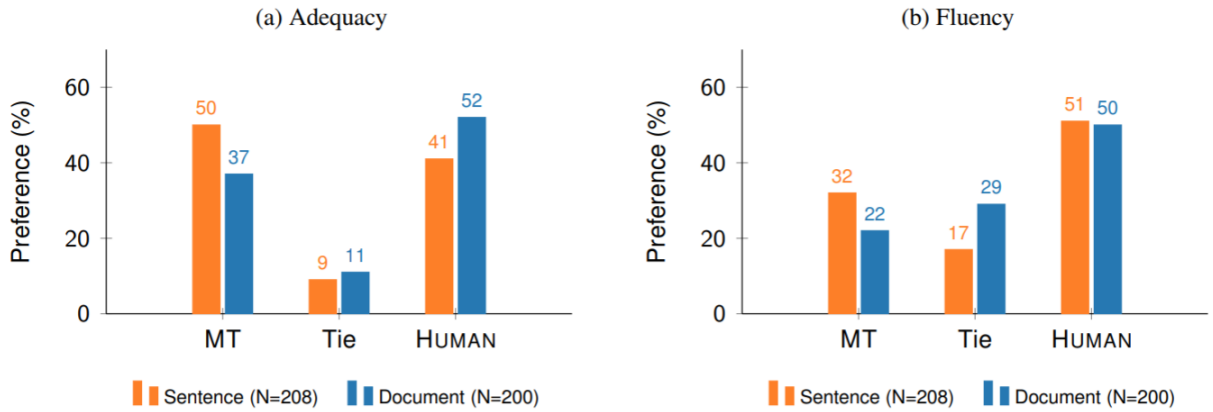


Document-level Transformer

Context-Aware Neural Machine Translation Learns Anaphora Resolution

Elena Voita, Pavel Serdyukov, Rico Sennrich, Ivan Titov

Where are we now?



arxiv.org/abs/1808.07048 : Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation

- Adequacy – evaluated by human translators
- Fluency – evaluated by users of language

- Sentence level adequacy is better for NMT than for human translators

Sentence level - problems

Anaphora

Lady Liberty geht voran. (….) Sie soll die Fackel der Freiheit von den Vereinigten Staaten in den Rest der Welt tragen.

Human

Lady Liberty is stepping forward. And just as Lazarus changed the meaning …. rightwingers have long wanted to change it back. She is meant to be carrying the torch of liberty from the United States to the rest of the world.

MT system (doc-level state of the art)

Lady Liberty is stepping forward. And just as Lazarus changed the meaning …. rightwingers have long wanted to change it back. It is meant to be carrying the torch of liberty from the United States to the rest of the world.

Sentence level - problems

Examples from Top WMT18 Systems

lexical coherence

Weidezaunprojekt ist elementar

Das Fischerbacher Weidezaun-Projekt ist ein Erfolgsprojekt und wird im kommenden Jahr fortgesetzt.

HUMAN	MT
Pasture fence project is fundamental	Electric fence project is basic
The Fischerbach pasture fence project is a successful project and will be continued next year.	The Fischerbacher Weidezaun-Project is a success and will be continued in the coming year.

Sentence level - problems

Coherence

99 documents reflect Austria's eventful history. Wolfgang Maderthaner, director-general of the Austrian State Archives, takes a very different approach.

99 documents reflect Austria's eventful history. This is not the case with Wolfgang Maderthaner, director-general of the Austrian state-owned army.

Testing Challenge

- Discourse phenomena don't show up in some established datasets
  - E.g. United Nations Corpus
- No good evaluation
  - BLEU? Not a good measure for discourse phenomena
    - It's fast and easy to calculate, especially compared to having human translators rate model output.
    - It's ubiquitous. This makes it easy to compare your model to benchmarks on the same task.

    BUT....
    - It doesn't consider meaning
    - It doesn't directly consider sentence structure
    - It doesn't handle morphologically rich languages well
    - It doesn't map well to human judgements

      [['early', 'puberty', ':', 'growing', 'older', 'sooner']]
      ['premature', 'puberty', ':', 'aging', 'earlier']
      BLEU: 6.867731683891005e-155

- We have new elaborate architectures… but are they really improving document level performance?
- All authors are testing on BLEU

BLEU algorithm

Ref: I ate three hazelnuts
Cand: Ate hazelnuts I three

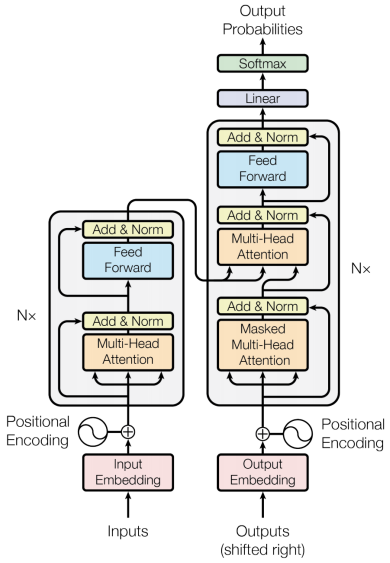
(n-grams)
[Ate hazelnuts]
[hazelnuts I]
[I three]

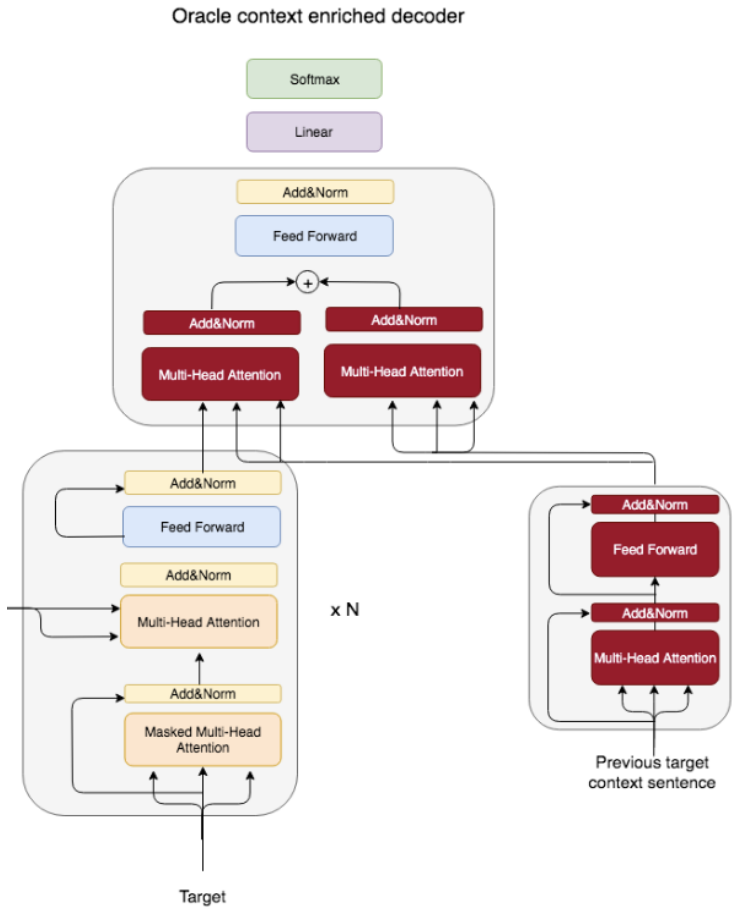
Language-model based measures are better. Always look for correlations of a measure with human annotators!
E.g. BERT score
https://github.com/Tiiiger/bert_score

It is freezing today and *The weather is cold today* will receive a high score.

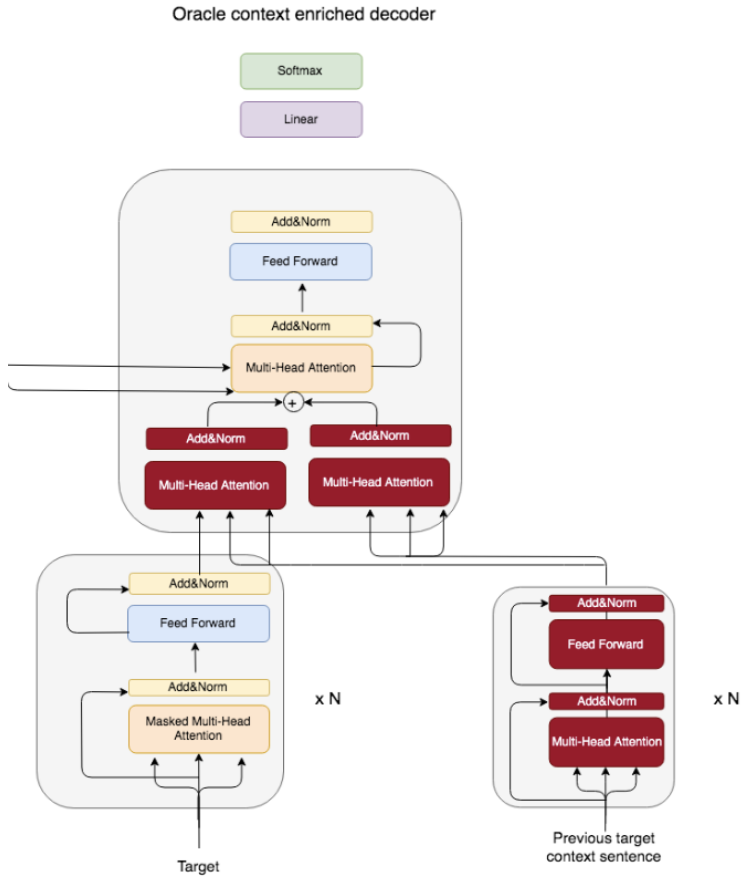
Next steps – new models

- 2 new models which incorporate target side context
  - Gated-Context: greater role of target context
  - Post-Cross: greater role of source context
- Interpretability
- Target context could be more useful than source context for some discourse phenomena





Gated-Context



Post-Cross

Adversarial Examples

Small perturbations to text induce dramatically different responses from models - "hacking" NLP models.

Error generation allows to test our solutions for their resiliency versus such hacking attempts.
https://github.com/MI2DataLab/WildNLP/tree/master
over 10 000 users

Figure 2: Textual QA [Ribeiro et al., 2018]: adversarial examples at work - replacement of words for their equivalents or words with minor spelling errors provokes the system to select another answer from source context.

In the United States especially, several high-profile cases such as Debra LaFave, Pamela Rogers, and Mary Kay Letourneau have caused increased scrutiny on teacher misconduct.
(a) Input Paragraph

Q: What has been the result of this publicity?
A: increased scrutiny on teacher misconduct
(b) Original Question and Answer

Q: What haL been the result of this publicity?
A: teacher misconduct
(c) Adversarial Q & A (Ebrahimi et al., 2018)

Q: What's been the result of this publicity?
A: teacher misconduct
(d) **Semantically Equivalent Adversary**

Figure 3: Visual QA [Ribeiro et al., 2018]: adversarial examples at work - reformulating a question provokes the system to label the image in an incorrect way.

What color is the tray?	**Pink**
What colour is the tray?	Green
Which color is the tray?	Green
What color is it?	Green
How color is tray?	Green

Table 1: Examples of text corruptions introduced by WildNLP aspects.

Aspect	Example sentence
Original	Warsaw was believed to be one of the most beautiful cities in the world.
`Article`	Warsaw was believed to be one of **a** most beautiful cities in world.
`Swap`	Warsaw **aws** believed to be one **fo teh** most beautiful cities in the world.
`Qwerty`	Wa**d**saw was b**d**lieved to be one of the most beautiful citi**e**e in the world..
`Remove_char`	Warsaw was believed to be one **o th** most **eautiful** cities in the world.
`Remove_space`	Warsaw was believed **tobe** one of the most beautiful cities in the world.
Original	You cannot accidentally commit vandalism. Vandalism used to be a rare occurrence.
`Misspelling`	You can not **accidentaly** commit vandalism. Vandalism used to be a rare **occurrance**.
Original	Bus Stops for Route 6, 6.1
`Digits2words`	Bus Stops for Route **six**, **six point one**
Original	Choosing between affect and effect can be scary.
`Homophones`	Choosing between **effect** and effect can **bee** scary.
Original	Laughably foolish or false: an absurd explanation.
`Negatives`	**Laughab*y fo*lish** or **fal*e**: an **a*surd** explanation.
Original	Sometimes it is good to be first, and sometimes it is good to be last.
`Positives`	Sometimes it is **go*d** to be first, and sometimes it is **goo*** to be last.
`Marks`	Sometimes, it is good to be first and sometimes**,** it**,** is good to be last.

Adversarial Examples

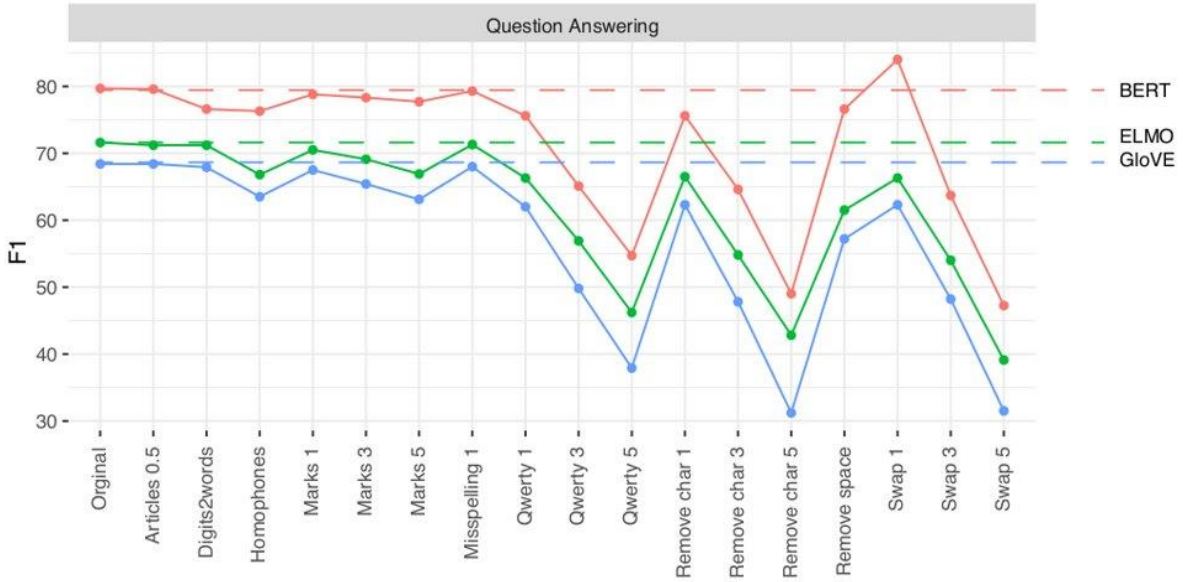


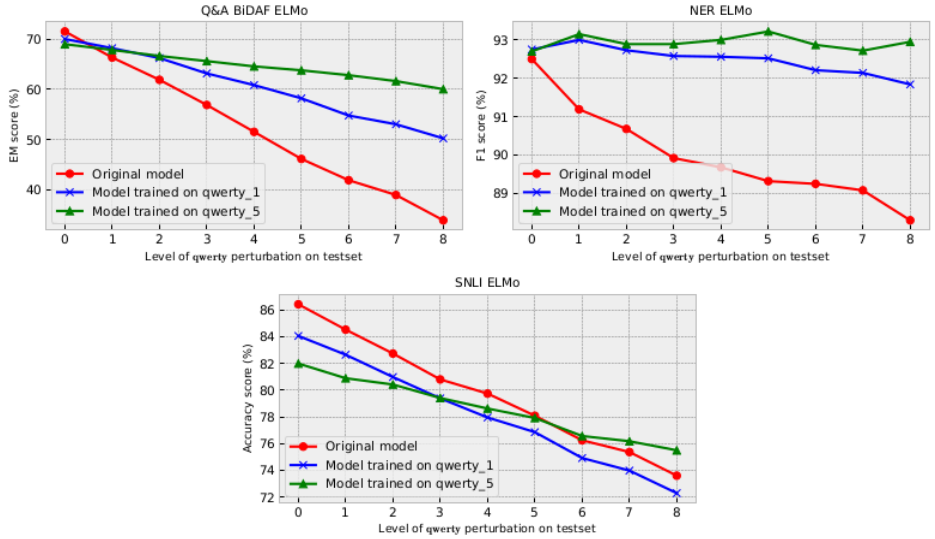
Figure 1: Robustness testing results for Q&A models.

src	1901 wurde eine Frau namens Auguste in eine medizinische Anstalt in Frankfurt gebracht.
adv	1901 wurde eine Frau namens **Afuiguste** in eine medizinische Anstalt in Frankfurt gebracht.
src-output	In 1931, a woman named **Augustine** was brought into a medical institution in France.
adv-output	In 1931, a woman named Rutgers was brought into a medical institution in France.
src	Das ist Dr. Bob Childs – er ist Geigenbauer und Psychotherapeut.
adv	Das ist Dr. Bob Childs – er ist Geigenbauer und **Psy6hothearpeiut**.
src-output	This is Dr. Bob Childs – he's a wizard maker and a **therapist**'s **therapist**.
adv-output	This is Dr. Bob Childs – he's a brick maker and a **psychopath**.

On Adversarial Examples for Character-Level Neural Machine Translation
Javid Ebrahimi, Daniel Lowd, Dejing Dou

Performance of Q&A models drops severely even if only slight modifications are introduced to questions...

Partial Solution?
Adversarial training on perturbed samples.



The article (for now anonymized):
https://openreview.net/pdf?id=SkxgBPr3iN

What's the lesson?

- Be a conscious user/designer of NMT systems
  - It's extremely useful
  - But has its (big) problems

- Evaluate, evaluate, evaluate
  - Test robustness
  - Challenge established datasets
  - Challenge measures
    - Do not use BLEU as first and only choice
  - Challenge established way of thinking about performance evaluation
    - Is a change of 0.001 in a metric really important?
    - What are worst-case scenarios for a model? Maybe it's better generally but it's worse-case performance is unacceptable.

As a last word...

I'm inviting everyone who finds these topics interesting to cooperate

- Within my grant titled MAD-NLP: Multiaspectual Diagnostics of NLP Systems (robustness)

- Within cooperation with NTU Singapore (NMT and other topics)

Thank you

Barbara Rychalska
b.rychalska@mini.pw.edu.pl
24.11.2019

Politechnika Warszawska

FINDWISE
SEARCH DRIVEN SOLUTIONS

NANYANG TECHNOLOGICAL UNIVERSITY

MI DATA LAB