University of Zurich[UZH]

Institute of Computational Linguistics

There is no experience in the use of ALLAH

Mathias Müller

And a more serious title: Domain Robustness in Neural Machine Translation

Mathias Müller

Collaborators

Thank you!

In a nutshell

- NMT models are bad at translating text from a domain they were not trained on
- We observe that a major symptom is **hallucination**: translations that are fluent, but unrelated to the source text
- We empirically test several strategies to increase domain robustness

Notion of domains

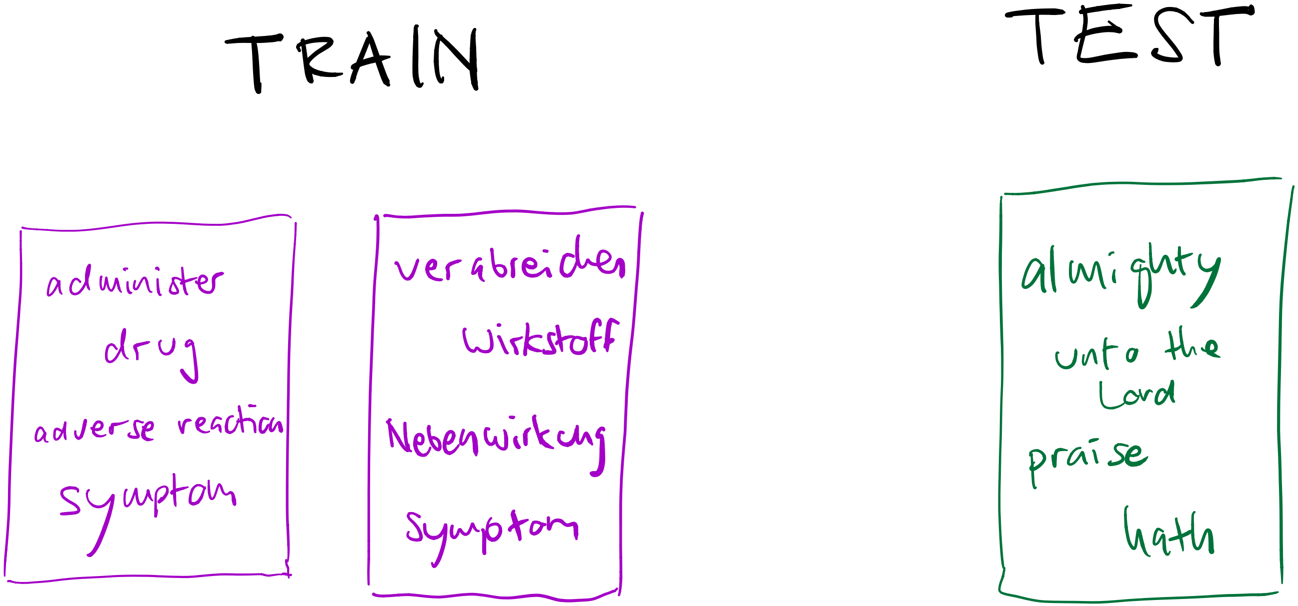
administer
drug
adverse reaction
symptom

medical

almighty
unto the
Lord
praise
hath

bible

Out-of-domain translation



(this should actually work!)

Domain Robustness

- property of being invariant to domain shift
- expresses actual goal of MT: to learn to translate in general, independent of domains
- different from other uses of the word "robustness": typos, adversarial attacks

Observation by Koehn and Knowles in 2017

train

BLEU

- NMT models cannot cope with domain shift

	NMT	SMT
Medical	39.4	43.5

test

Law	3.9	10.2
IT	2.0	8.5
Koran	0.6	2.0
Subtitles	1.4	5.8
Average	**2.0**	**6.6**

Comparison to SMT

- important realization: SMT systems have higher domain robustness, sometimes drastically
- proves that it is possible to generalize better to unseen domains

Questions

- Outcome the same for deeper models or different architectures?

Deep RNN   Transformer

- If yes, how can we improve the domain robustness of NMT?

With current models

- Same outcome with deeper models?

	Koehn and Knowles		Ours	
	RNN	SMT	Deep RNN	SMT
Medical	39.4	43.5	57.5	58.4
Law	3.9	10.2	17.4	19.8
IT	2.0	8.5	11.6	21.4
Koran	0.6	2.0	1.1	1.4
Subtitles	1.4	5.8	1.6	4.7
Average	**2**	**6.6**	**8.7**	**11.8**

Noticed: Odd translations

Source sentence from subtitles domain

Aber geh subtil dabei vor.

Target sentence (reference)

But be subtle about it.

Nematus RNN Baseline trained on medical domain

Pharmacokinetic parameters are not significantly affected in patients with renal impairment (see section 5.2).

Hallucination

- Even in-domain, NMT models occasionally fall into a hallucination mode

- Is hallucination more prominent in out-of-domain translation?

Manual analyses of fluency and adequacy

Is this target sentence a translation of the source sentence?

ADEQUACY

Is this target sentence fluent, grammatical English?

FLUENCY

Manual analysis of adequacy

	adequate	partially	inadequate
Medical	54	44	2
Law	14	60	26
IT	11	48	41
Koran	0	25	75
Subtitles	3	22	75

#Sentences = 600. All numbers are in %.

Manual analysis of fluency

only

Fluency of **inadequate** translations:

fluent	partially	not fluent
44%	19%	37%

Out-of-domain translation: another example

Source sentence from Koran domain

Dann fanden sie für sich anstelle von ALLAH keine Beistehende.

Sockeye RNN Baseline trained on medical domain

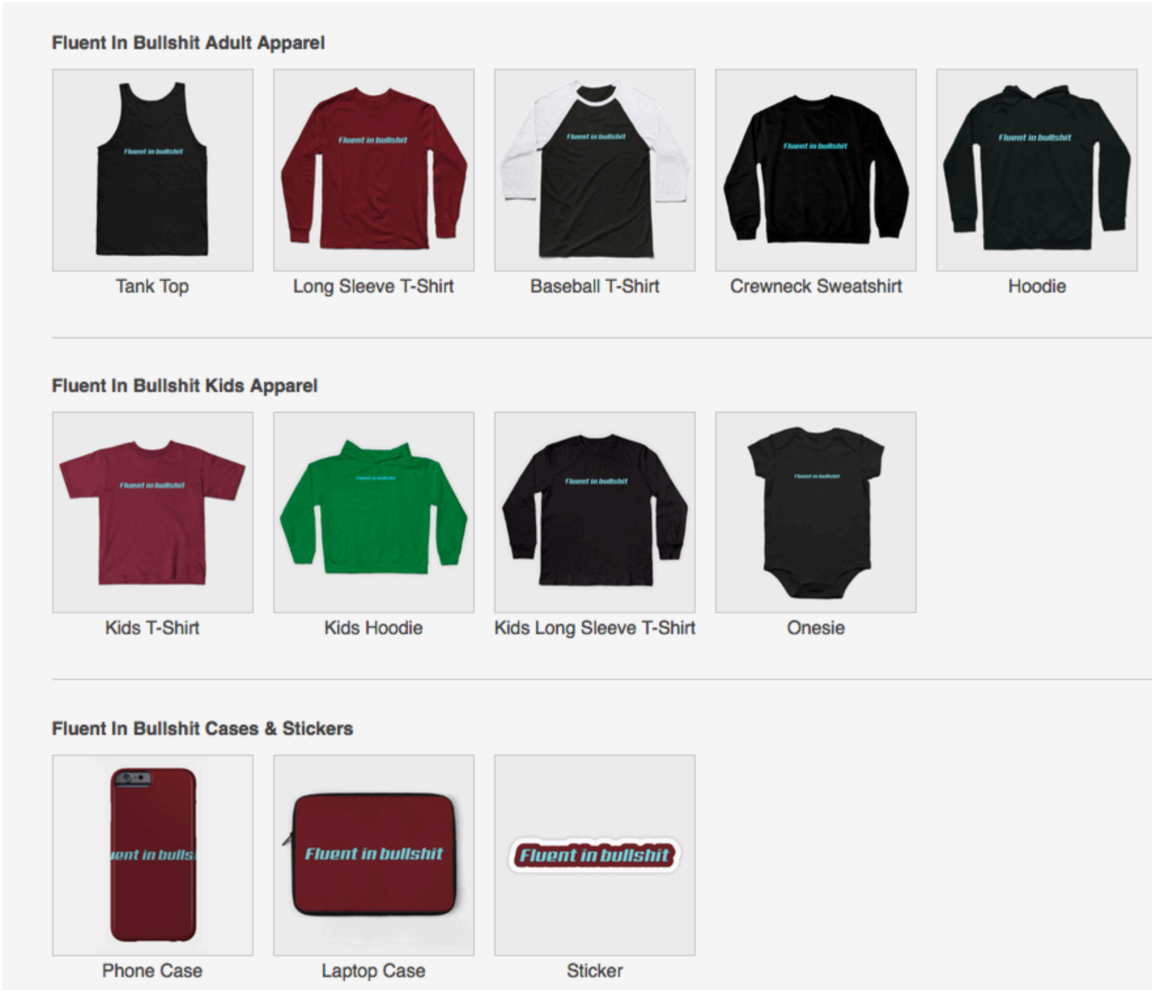
There is no experience in the use of ALLAH.

mtrain SMT Baseline trained on medical domain

Then it for you instead of ALLAH no Beistehende.

This could be a T-shirt!

It already is a T-shirt :(



Strategies to mitigate the problem

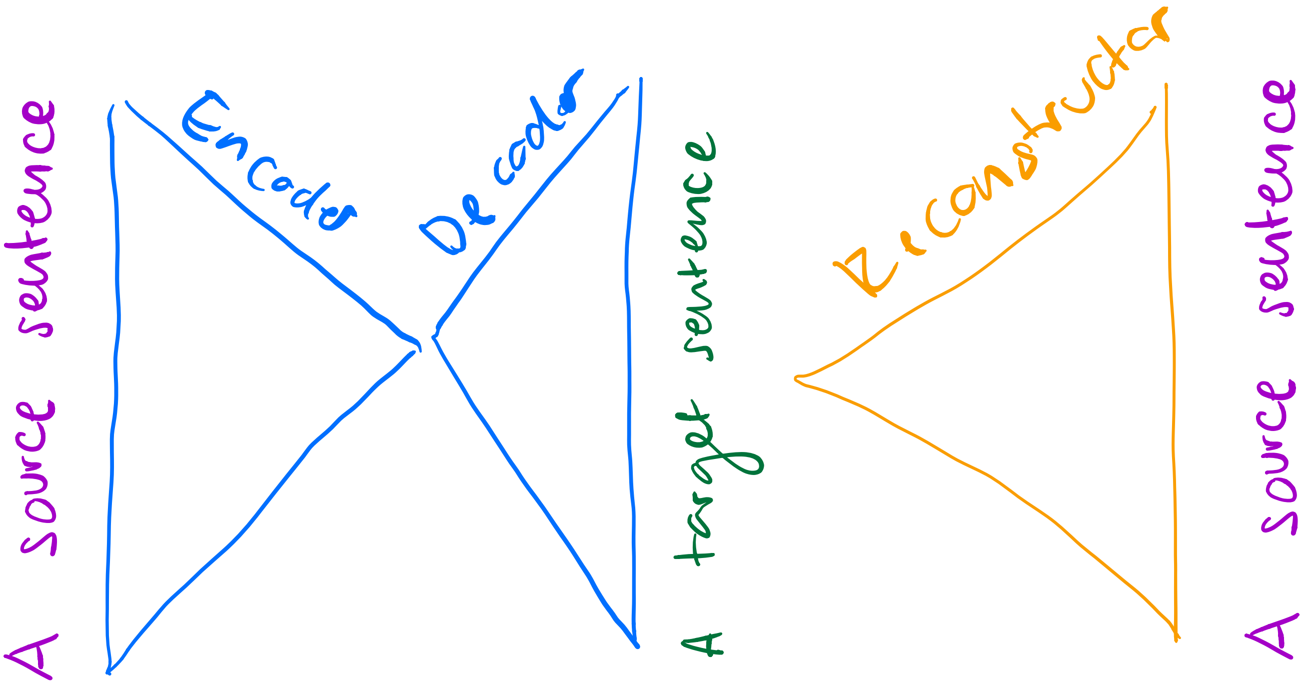
- Reconstruction
- Defensive distillation

MIGROS
SÉLECTION

2/4

Reconstruction

An additional network must be able to translate from the decoder states to the original source sentence

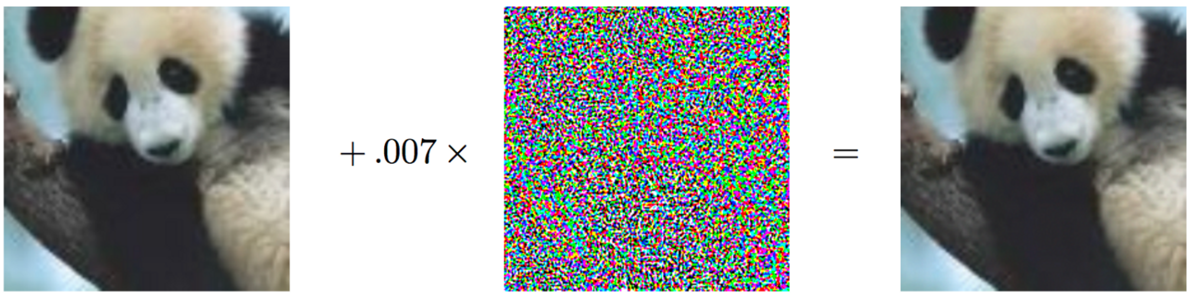


Reconstruction Results

	Baselines		
	RNN	**SMT**	**Reconstruction**
Medical	57.5	58.4	58.4
Law	17.4	19.8	20.4
IT	11.6	21.4	17.5
Koran	1.1	1.4	1.1
Subtitles	1.6	4.7	2.9
Average	**8.7**	**11.8**	**10.4**

Defensive Distillation

- use distillation to guard against adversarial attacks



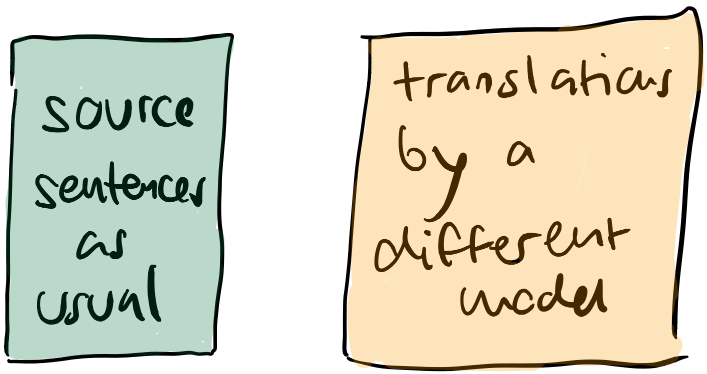
98% panda

99% gibbon

Defensive Distillation

- Distillation: learn a model (**student**) on predictions of another model (**teacher**) instead of gold labels

train on :



- distillation was found to harden networks against adversarial attacks

Defensive Distillation for Domain Robustness?

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \qquad \begin{bmatrix} 2 \\ 4 \\ 16 \\ 7 \end{bmatrix}$$

out-of-domain translation ≈ adversarial examples

$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \qquad \begin{bmatrix} 1.999 \\ 4.012 \\ 16.008 \\ 6.986 \end{bmatrix}$$

Defensive Distillation Results

	Baselines		
	Transformer	SMT	Defensive Distillation
Medical	61.2	58.4	60.7
Law	20.2	19.8	21.1
IT	13.8	21.4	15.6
Koran	0.8	1.4	0.9
Subtitles	1.9	4.7	2.9
Average	**9.2**	**11.8**	**10.1**

Defensive Distillation Discussion

- Results indicate a relationship between adversarial examples and out-of-domain translation

Summary

- NMT models exhibit low domain robustness
- Symptom: hallucination is more pronounced in out-of-domain translation
- Among the strategies we tested,
  - **reconstruction** gave the best results
  - **distillation** the most intriguing results

University of Zurich[UZH]

Institute of Computational Linguistics

Backup Slides

Next steps

- Our training, dev and test sets will be freely available (in fact, already given to other people from UEDIN)
- Work submitted to ACL 2019

Future work

- Strengthen currently weak argument for defensive distillation
- theoretical measures of domain distance, e.g. **A-distance**

With current models

- Same outcome with Transformer models?

	Our Models		
	RNN	**Transformer**	**SMT**
Medical	57.5	61.2	58.4
Law	17.4	20.2	19.8
IT	11.6	13.8	21.4
Koran	1.1	0.8	1.4
Subtitles	1.6	1.9	4.7
Average	**8.7**	**9.2**	**11.8**