Towards the Creation of Machine Translation Systems Between Russian and Turkic Languages “TurkLang-7’

A. Khusainov,
A. Gatiatullin, D. Suleymanov, R. Gilmullin
Institute of Applied Semiotics TAS

Main presentation topics

- The main idea of this project
- What we’ve already done
- What’s our plans
- Conclusions

TurkLang - 7

Goal:

- to provide Russian-Turkic language pairs with high-quality machine translation

Tasks:

- (obviously) creating datasets
- (obviously) experimenting with tools/algorithms
- (obviously) building models
- creating web-site (not demo, not experimental, real-working)
- (maybe more important) unite the efforts of people

* Masakhane project: MT for African languages

TurkLang - 7

Goal:

- to provide Russian-Turkic language pairs with high-quality machine translation

Why it’s important:

- preservation and development of the language
- active use of the language in the Internet
- the possibility of high-quality translation of documents
- communication and study of languages

TurkLang - 7

Goal:

- to provide Russian-Turkic language pairs with high-quality machine translation

◆ **Tatar** ◆ **Bashkir** ◆ **Kazakh** ◆ **Chuvash** ◆ **Uzbek** ◆ **Kirgiz** ◆ **Crimean-Tatar**

58 million speakers*

*According to Ethnologue project
https://www.ethnologue.com/

Standard MT recipe

DATA

TECHNOLOGY

HARDWARE

we can get good results more or less straightforward when

we have ALL those things

TurkLang – 7. Institute’s background

Data:

- Own corpus-manager tools (tugantel.tatar)
- First speech corpora for Tatar ASR and TTS
- Parallel corpus for Russian-Tatar pair

Tech:

- Statistical-based MT
- Neural MT: encoder-decoder-attention
- Neural MT: Transformer-based

Hardware:

- 12x Tesla V-100 32GB GPU cards

Main Stages

of the TurkLang-7 project

TurkLang – 7. Stages

Data collection process and technologies:

- Collect existing corpora
- Make a list of bilingual sources (esp. web-sites)
- Download info, boilerplate removal
- Document-level alignment via MT / lexicon comparison
- Segment-level alignment via MT / lexicon comparison
- Sentence splitter, de-duplicate, filter

- + monolingual data + other linguistic resources
- + manual work

TurkLang – 7. Stages

Algorithms:

- Transformer neural network model
- Ensembles
- NN LM fusion
- Back-translation approach
- Fine-tuning approaches
- Sub-word segmentation
- Various NN search algorithms
- (!) Using rule-based methods based on Turkic Morpheme Model

TurkLang – 7. Stages

Web-site:

- Python-based web-server
- NN inference server
- Server balancing
- Using both GPU and CPU
- Teacher-student approach
- User’s feedback
- Multilingual localization

Preliminary Results

TurkLang – 7. Results so far

- Community building
- We established semi-automatic process of data collection
- Developed necessary software for sub-tasks
- Run several experiments on already existing corpora

Plans for:

- rule-based data augmentation and series of NN training experiments
- web-site creation and (stress) testing

TurkLang – 7. Numbers

Language pair	# of sources	# of parallel sentences
Tatar-Russian	+3	+439 000
Bashkir-R.	8	388 000
Chuvash-R.	1	206 000
Kirgiz-R.	9	471 000
Uzbek-R.	-	-
Kazakh-R.	1	5 000 000
Crimean-Tatar-R.	-	-

4. Tatar-Russian MT system

Architecture and experiments

Characteristics of Tatar-Russian parallel corpus

Parameter	Value
# of parallel sentences	**983 319**
# of words in Russian part	15 032 363 (15,3 words per sentence)
# of words in Tatar part	14 649 484 (14,9 words per sentence)
# of sentences in train/test/valid	977539 / 2499 / 2499

Methods and NN types

NN size

Transformer Base: batch size – 2048, hidden size – 512, filter size – 2048, multi-headed attention heads – 8, encoder/decoder's hidden layers – 6, dropout – 0.1, learning rate – 2.0, beam size – 4.

Transformer Big: x2 batch size, hidden size, filter size, multi-headed attention heads

Back-translation

No BT corpus

+0.5x of “real” parallel

+1x of “real” parallel

Transfer learning

Data from WMT 2019

Basic approach of pre-training (without layers' freezing)

LM integration

Deep fusion:

LM+TM

Weighted sum of LM+TM

Results. Base/Big

Model type	Iteration count	Translation direction	BLEU
Base	10	RU-TT	33.57
Base	20	RU-TT	34.82
Base	30	RU-TT	35.27
Base	**40**	**RU-TT**	**35.39**
Big	10	RU-TT	34.08
Base	10	TT-RU	35.95
Base	20	TT-RU	37.71
Base	30	TT-RU	38.41
Base	**40**	**TT-RU**	**38.42**
Big	10	TT-RU	37.07

Results. Back-translation + search algorithm

Model type + iteration count	Search algorithm	Translation direction	BLEU
0.5x 10	beam	TT-RU	36.84
0.5x 20	beam	TT-RU	37.73
0.5x 30	beam	TT-RU	38.50
0.5x 40	beam	TT-RU	38.63
0.5x 40	**beam**	**RU-TT**	**34.89**
1x 40	**beam**	**TT-RU**	**39.21**
1x 40	beam	RU-TT	34.42
1x 40	random	TT-RU	18.21
1x 40	random	RU-TT	17.73

Results. Transfer learning

Model type	Iteration count	Translation direction	BLEU
Base	10	RU-KK	50.01
Base	+ 10	RU-KK-TT	34.41
Base	10	KK-RU	61.47
Base	+ 10	TT-KK-RU	36.08

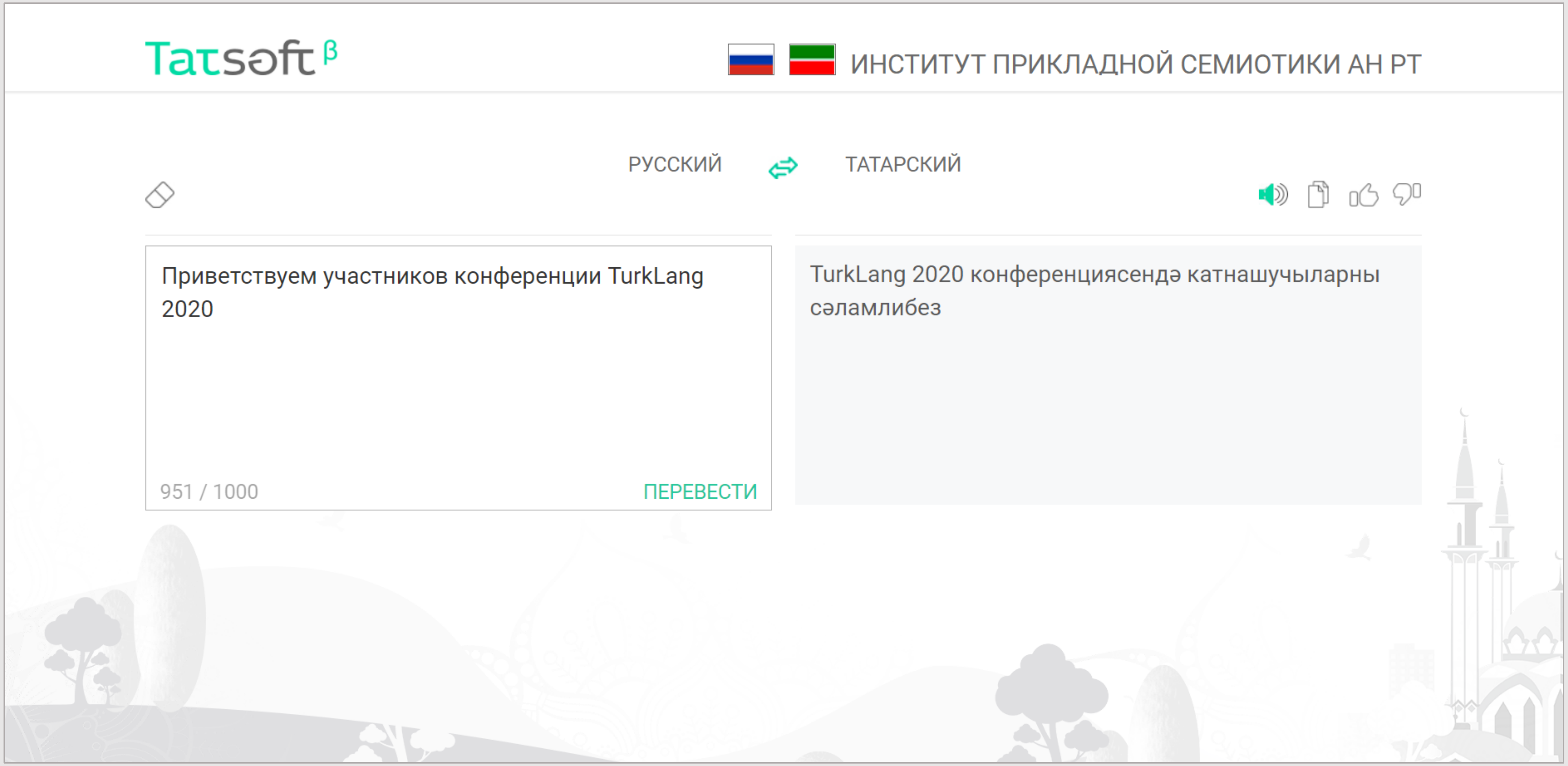
Results. LM deep fusion

Model type	Iteration count	Translation direction	BLEU
Sum of logits of LM and TM	30	RU-TT	32.73
Weighted sum of logits of LM and TM (a=0.1)	20	RU-TT	34.56

Results. Overview and comparison

Model type	Translation direction	BLEU
Base 40 w/o LM, BT, TL	RU-TT	35.39
Yandex	RU-TT	15.59
Google	RU-TT	17.00
Base 40 + 1x beam-search BT	TT-RU	39.21
Yandex	TT-RU	18.16
Google	TT-RU	22.64

translate.tatar



Приветствуем участников конференции TurkLang 2020

951 / 1000

TurkLang 2020 конференциясендә катнашучыларны сәламлибез

Thanks!

Any questions?

You can find me at:

khusainov.aidar@gmail.com