Tagging and editing of the Russian-Tatar parallel MT testing corpus

Bulat Khakimov, Marat Shaekhov

Applied Semiotics Research Institute
Tatarstan Academy of Sciences

TurkLang-2021, 22.09.2021. Kyzyl

Public NMT systems for Tatar-Russian language pair available online:

Yandex (2019)

Tatsoft (2019)

Google (2020)

Promt.One (2020)

Tatsoft (http://translate.tatar)

- publicly available machine translation service.
- developed by Applied Semiotics Institute in Kazan, Tatarstan (Russia)
- based on neural networks and machine learning
- speech recognition and synthesis
- special models for the Tatar language

MT evaluation methods

- automated/quantitative
  - *BLEU (Papineni et al, 2002)*
  - *Post-editing distance (Levenshtein, 1966)*
- expert/qualitative
  - *equivalence*
  - *linguistic features*
  - *etc.*

BLEU test of TT-RU MT engines

1000 random sentences:

- professional human translations
- Yandex translations
- Tatsoft translations
- Google translations

BLEU metric

MT engine	RU-TT	TT-RU
Yandex	15.59	18.16
Tatsoft	**35.39**	**39.21**
Google	17.00	22.64

We need a more accurate and detailed translation quality evaluation methodology.

The results of the quantitative tests like BLEU may be overvalued and do not reflect the actual quality of the translation, especially when using a test sample from the same sources as the training data.

Combined approach

- linguistic analysis of typical errors in texts of different styles and topics
- identification of linguistic markers for contexts with typical errors
- quantitative assessment of the frequency of contexts with different markers (including POS tags and their combinations)

Main concepts of MT testing corpus collection

- diversity
- stylistic representativeness
- linguistic representativeness

Data collection

- ✔ 2184 parallel sentences
- ✔ 27613 words in Tatar and 28361 words in Russian
- ✔ manual human translations
- ✔ different published sources
- ✔ mostly RU-TT

News, mass media

Сегодня в Национальном культурном центре "Казань" состоялся праздничный концерт, приуроченный ко Дню работника прокуратуры России.	Бүген "Казан" милли мәдәният үзәгендә Россия прокуратурасы хезмәткәрләренең һөнәри бәйрәмнәре уңаеннан бәйрәм концерты булды.
Не прошло и 50 лет, как к сноуборду появился интерес и в России.	50 ел да узмады, сноуборд белән кызыксыну Россиядә дә барлыкка килде.

Fiction

Подозревали Напарника, которому Дикий старался побольше насолить.	Дикий күбрәк зыян күрсәтергә тырышкан Напарниктан шикләнделәр.
Газинур бежал вместе со всеми.	Газинур башкалар белән бергә йөгерде.

Official Documentation

Процедуры, устанавливаемые настоящим пунктом, осуществляются в день прибытия заявителя.	Шушы пункттан чыгып билгеләнә торган процедуралар мөрәҗәгать итүче килгән көнне гамәлгә ашырыла.
О проекте федерального закона № 607441-5 "О внесении изменений в Федеральный закон "О физической культуре и спорте в Российской Федерации" (в части совершенствования деятельности спортивных федераций).	Россия Федерациясендә физик культура һәм спорт турындагы Федераль законга үзгәрешләр кертү хакында" 607441-5 номерлы федераль закон проекты турында (спорт федерацияләре эшчәнлеген камилләштерү өлешендә).

Scientific and educational literature

Важнейшая часть клетки - ядро.	Күзәнәкнең иң әһәмиятле өлеше - төш.
Если в пробирку с раствором щелочи прибавить несколько капель раствора медного купороса и к образовавшемуся гидроксиду меди (II) прилить глицерин, то образуется прозрачный раствор глицерата меди ярко-синего цвета (из-за сложности строения получающегося вещества формула его не приводится).	Әгәр селте эремәсе салынган пробиркага берничә тамчы бакыр купоросы эремәсе тамызып, хасил булган бакыр (II) гидроксидына глицерин салсаң, ачык зәңгәр төстәге үтә күренмәле бакыр глицераты эремәсе барлыкка килә (барлыкка килгән матдәнең төзелеше катлаулы булу сәбәпле, формуласы китерелми).

Religious texts

Спустя несколько времени, Каин принёс от плодов земли дар Господу.	Берникадәр вакыт узгач, Кабил җир җимешләре уңышыннан Раббыга бүләк китерде.
И исполнил его Духом Божиим, мудростью, разумением, ведением и всяким искусством.	һәм аны, Аллаһы Рухы белән рухландырып, барлык һөнәрләр өчен кирәкле акыл, осталык һәм төрле күнекмәләр белән баеттым.

Spoken language

Подлая ложь, товарищи!	Чеп-чи ялган, иптәшләр!
Вот то-то и оно-то!	Менә шул-шул!

Editing and tagging

- ✔ collection of parallel sentences
- ✔ spell check and correction
- ✔ punctuation normalization
- ✔ abbreviations, special symbols, numbers normalization
- ✔ manual equivalence check and adaptation
- ✔ automated POS tagging
- ✔ manual disambiguation using “Tugan tel” corpus manager (http://tugantel.tatar)
- ✔ automated and manual mark-up of the sentences

Linguistic features

- ✔ marked automatically
- ✔ marked manually
- ✔ RU
- ✔ TT

Conclusion and Perspectives

The parallel testing corpus, balanced by styles and linguistic features, allows to perform a more accurate and relevant quality evaluation of the machine translator with both quantitative and qualitative methods.

In turn, this makes it possible to “fine-tune” the training sample in conditions of limited bilingual data.

Current work includes the extension of the text corpus, study of frequency and typical errors in the test sample, study of significant style markers in the Tatar language.

Игътибарыгыз өчен рәхмәт!