PROMT®

Creating an Automated System for Translation of User-Generated Content

Plan

1. Parties
2. Problem
3. Solution
4. Translation Quality Evaluation
5. Conclusions

About PROMT

Experienced.

More than 20 years of experience in developing machine translation technology and products for server, desktop and mobile.

Diversified.

60 language pairs and over 180 domain-specific dictionaries.

Widely used.

Over 100 million hits per month on our online translation sites (www.translate.ru and www.online-translator.com)

About TripAdvisor

1. World’s largest travel site (over 60 M unique monthly visitors)
2. Enables travelers to plan perfect trips
3. Offers trusted advice and reviews from real travelers (over 75 M reviews)
4. Most travel reviews presented in English
5. Users want to read reviews in their native language

About Project

1. Fast growth of Russian travel market
2. Russian version of TripAdvisor’s website
3. All reviews in English should be translated into Russian
4. Human translation cannot be implemented due to huge amounts of content
5. An efficient machine translation solution is needed to solve the task

Requirements to MT provider

- High translation quality sufficient for understanding without human post-editing
- Technically accurate server-based solution for processing large volumes of text data in real time
- Integration into the workflow of TripAdvisor’s website
- Automatic quality estimation technology integrated into the machine translation solution

Challenges

User-generated content (UGC)

- Similarity to oral content
- Spelling errors
- Grammar and syntax errors
- Style of writing determined by cultural, linguistic, emotional features of authors

Data provided for training

- Not enough parallel in-domain data for training the English-Russian engine
- No target in-domain data provided

Tight deadline for developing MT solution

PROMT solution components

PROMT Translation Server 9.5 DE

a reliable, robust, and scalable server-based solution that allows the translation of large text volumes in real time

PROMT DeepHybrid Technology

high-quality MT comprehensible for end users

Quality Estimation Technology

PROMT DeepHybrid Technology

PROMT Baseline Translation Engine

Statistical Component

Linguistic Databases:
Dictionaries,
Translation Memories

Language Model

Creating Dictionaries for PROMT Solution

1. Conversion of the TripAdvisor English-Russian glossary into a dictionary of the PROMT internal format

⇨ **Dictionary with client-specific terms (≈5,600 entries)**

2. Extraction of the most frequent domain-specific terms and phrases from the English hotel reviews corpus

3. Creation of a dictionary with incorrect spelling of frequent English words

⇨ **Background dictionary with domain-relevant terminology (≈27,600 entries)**

4. Creation of a dictionary containing geographic names using PROMT internal resources

⇨ **Geography background dictionary (≈48,200 entries)**

Result: an extensive linguistic in-domain database; ability to recognize and process errors, misprints, and abbreviations.

Translation Memories

English hotel review corpus (1.2 billion words)

- Building a list containing the most frequent in-domain sentences
- Translating these sentences with the PROMT baseline system
- Expert analysis of the translations
- Manual post-editing of bad translations

Result: an in-domain TM database integrated into the translation solution.

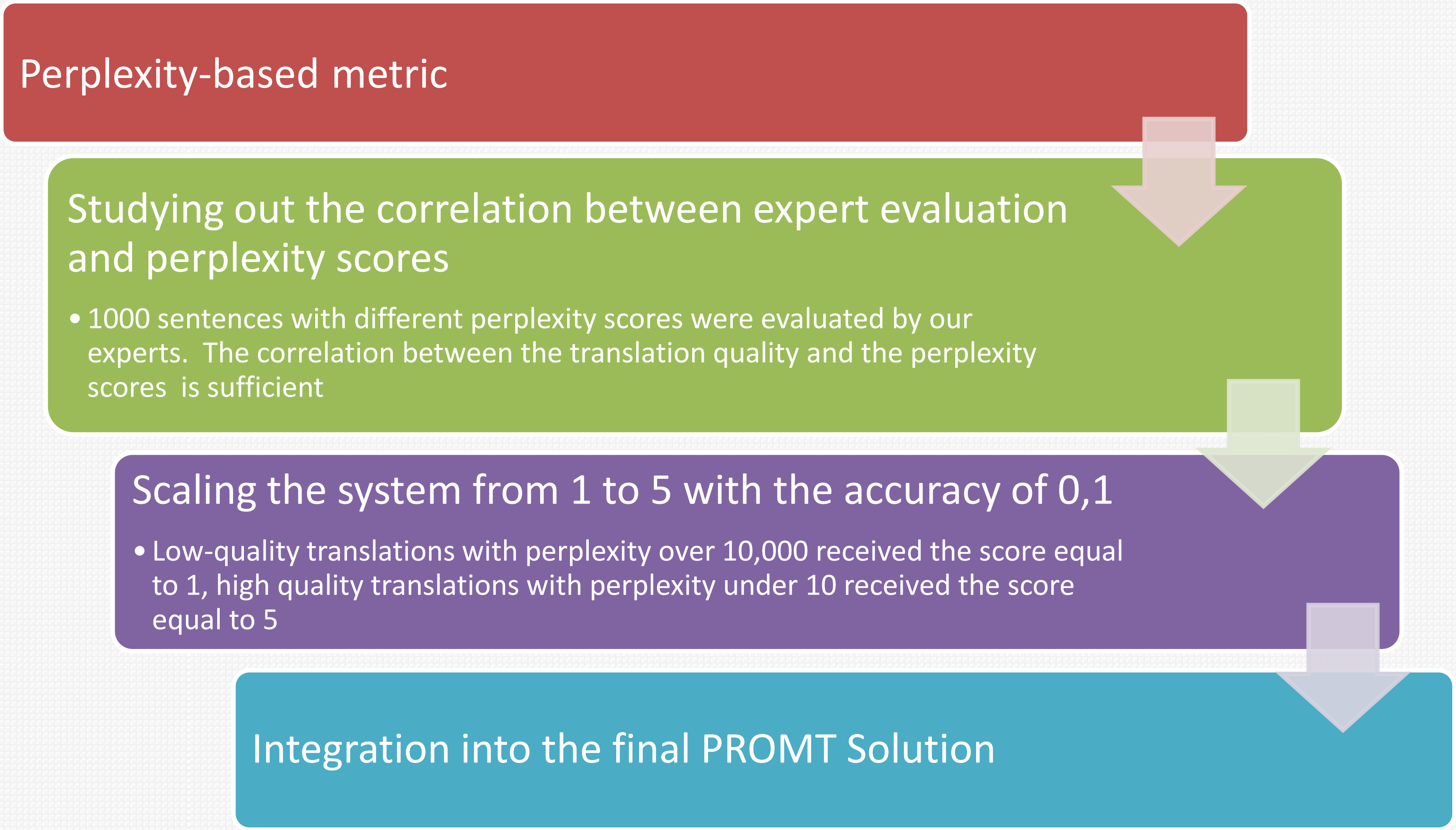
Creating the Language Model

No target in-domain data was provided by TripAdvisor

- Downloading about 27,000 user reviews (80 million words) from different Russian travel websites,
- Filtering the downloaded data, creating a target in-domain corpus,
- Building the target language model.

Result: a target language model integrated into the translation solution.

Quality Estimation System



Perplexity-based metric

Studying out the correlation between expert evaluation and perplexity scores

- 1000 sentences with different perplexity scores were evaluated by our experts. The correlation between the translation quality and the perplexity scores is sufficient

Scaling the system from 1 to 5 with the accuracy of 0,1

- Low-quality translations with perplexity over 10,000 received the score equal to 1, high quality translations with perplexity under 10 received the score equal to 5

Integration into the final PROMT Solution

Translation Quality Evaluation

Test set:

English-Russian Parallel corpus of reviews (≈ 70K words)

System	BLEU score	Percentage of unknown words	Expert evaluation (DeepHybrid compared to baseline)		
			Improvements	Degradations	Equal translations
PROMT baseline system	17.12	2,56%			
PROMT baseline system + TripAdvisor dictionaries	19.42	2,19%			
PROMT DeepHybrid system (PROMT baseline system + TripAdvisor dictionaries + Language model)	20.13	2,16%	49	9	42

Examples of translation quality improvements

№	Source sentence	PROMT Baseline System	PROMT DeepHybrid Technology	google.translate
1	A big thumbs up to the Kiydan family	Большие большие пальцы до семьи Kiydan	Оценка «отлично» семье Киидэн	Большие пальцы в семье Kiydan
2	Can’t wait to go back!!	Не может ждать, чтобы возвратиться!!	Не терпится вернуться снова!!	Не может ждать, чтобы вернуться!
3	The **brakfast** was awsome.	brakfast был awsome.	Завтрак был потрясающим.	Завтраком было потрясающим.
4	The food and **resturant** was very good	Еда и resturant были очень хороши	Еда и ресторан были очень хороши	Еда и ресторан был очень хорош
5	At least the staff were **pleasent**!	По крайней мере, сотрудники были pleasent!	По крайней мере, персонал был приятным!	По крайней мере, сотрудники были **приятно!**
6	Dinner at the hotel was quite expensive and we preferred to eat out, however we ate at the hotel one day when the **menue** included lobster.	Обед в отеле был довольно дорог, и мы предпочли идти куда-нибудь поесть, однако мы поели в отеле однажды, когда menue включал омара.	Ужин в отеле был довольно дорогим, и мы предпочли идти куда-нибудь поесть, однако мы поели в отеле однажды, когда меню включало омара.	Ужин в отеле был довольно дорогим, и мы предпочли пойти куда-нибудь поесть, но мы поели в отеле однажды, когда МЕНЮ включены **омаров**.

Output solution features:

- Technically accurate server-based solution for processing large volumes of text data in real time.
- High quality translation
- Low costs for development and customization of the MT solution (compared to the manual translation costs).
- Accurate and efficient quality estimation system.

The solution was integrated into the TripAdvisor workflow with minimal costs for development and support on the client’s side. The solution was deployed on a dedicated server in the PROMT datacenter.

Future work

Integration of statistical post-editing component into the translation solution:

1. Crawling and aligning comparable in-domain data from Web (travel websites)
2. Collecting the parallel data provided by TripAdvisor (most frequent reviews translated by the PROMT system and manually post-edited)
3. Using the accumulated data for training the statistical post-editing component

PROMT®

Alexander Molchanov
Alexander.Molchanov@promt.ru

Leonid Evdokimov

Leonid.Evdokimov@promt.ru

Linguistic Department
PROMT Ltd
Tel.: +7 (812) 611-00-50
www.promt.ru
www.translate.ru
Saint-Petersburg, Russia