ОПЫТ НАСТРОЙКИ СИСТЕМЫ АВТОМАТИЗИРОВАННОГО ПЕРЕВОДА ПОЛЬЗОВАТЕЛЬСКОГО КОНТЕНТА

Евдокимов Л. В. (Leonid.Evdokimov@promt.ru),
Молчанов А. П. (Alexander.Molchanov@promt.ru)

ООО «ПРОМТ», Санкт-Петербург, Россия

В данной статье описывается опыт компании PROMT по настройке и реализации системы автоматизированного перевода PROMT Deep-Hybrid для интерактивной обработки текстовой информации, представленной на сайте, который представляет собой крупный интернет-ресурс, посвященный туризму и путешествиям. Целью работы было создание решения для перевода пользовательского контента, состоящего из текстов отзывов об отелях, ресторанах и других составляющих современного туристического сектора.

Ключевые слова: машинный перевод, пользовательский контент, автоматизированный перевод, гибридная технология перевода

CREATING AN AUTOMATED SYSTEM FOR TRANSLATION OF USER-GENERATED CONTENT

Evdokimov L. V. (Leonid.Evdokimov@promt.ru),
Molchanov A. P. (Alexander.Molchanov@promt.ru)

PROMT Ltd., Saint-Petersburg, Russia

This paper describes fast implementation of a hybrid automated translation system for processing user-generated content. We report on engine customization for TripAdvisor, the world's largest travel website. Due to the growing potential of the Russian travel market, TripAdvisor created the Russian version of its website and decided to translate all English reviews into Russian. PROMT, a leading provider of industrial MT solutions, was selected as MT vendor for the English-Russian language pair. According to the client's request we had to perform customization within a short period.

All input data represent user-generated content, so we faced several problems while building a large-scale, robust, high-quality engine. We decided to create a solution based on a hybrid machine translation system for the hybrid approach makes possible fast and efficient customization of a translation system with little or none in-domain data.

We automatically crawled a large web-based Russian text corpus of tourist reviews to build a statistical language model for our hybrid translation

system. We analyzed a batch of tourist reviews in English provided by TripAdvisor, created a number of dictionaries, a translation memory and defined translation rules for user-generated content. To handle the problem of various typos and misspellings we added most frequent misspelled words and phrases to the created dictionaries.

We experimented on a test set of tourist reviews in English provided by TripAdvisor. We report on improvements over our baseline system output both by automatic evaluation metrics and linguistic expertise.

Keywords: machine translation, user-generated content, automated translation, hybrid technology

1. Introduction

The fast evolution of computers and the rapid growth of the Internet since the late 1990s made it easier for people to upload, store and share information on the web. Forums, chats and other web-based informational resources led to the emergence of large amounts of the so called 'user-generated content'. User generated content (UGC) is material on websites, and occasionally other media sources, that is produced by users of websites (who are generally amateurs as opposed to professional editors, copywriters etc). In our case the content consists of tourist reviews produced by the users of the tripadvisor.com website.

TripAdvisor is world's largest travel web-based resource. The content is available in 21 languages for 30 countries. Most reviews are presented in English. At the same time, millions of users want to read the reviews in their native language. Human translation cannot be efficient for processing large amounts of UGC. Taking into account the fast growth of Russian travel market, TripAdvisor wanted an efficient automated translation solution for processing UGC.

2. Related Work

Regardless of the growing demand for automated translation of UGC little attention is paid to this topic in the field of machine translation research.

[Flournoy and Callison-Burch, 2000] discuss the possibility of creating a high-quality commercially successful application for real-time automated translation of chat content. The authors note that UGC is characterized by specific repeated colloquial words and phrases and lots of grammar errors. The main task of an MT system is to convey the meaning of the source text, whereas the translation quality is of secondary importance.

[Flournoy and Rueppel, 2010] investigate the development of an automated translation system for Adobe. The authors define three types of UGC:

- user e-mails;
- bug reports and product reviews;
- messages from user forums.

According to the authors, an efficient MT system for processing UGC should have the following features:

- ability to translate large amounts of texts in real time;
- ability to convey source text meaning;
- reliability and robustness (taking into account large volumes and low quality of input data).

[Banjeree et al., 2011] and [Banjeree et al., 2012] present the case-studies of customization of an automated MT system for processing UGC from the Symantec company forum. Authors observe a lot of grammar mistakes and a large number of colloquial words and phrases in the analyzed texts.

[Jie Jiang et al., 2012] report on the customization of an automated translation system for user messages in a multilingual social network. The authors face the following problems:

- a lot of the content is produced by non-native speakers, therefore this content contains many grammatical and syntactic errors;
- the content produced by native speakers contains grammatical and syntactic errors because 1) either the author enters the text too fast and so makes typographical errors, or 2) the author deliberately departs from spelling norms to bring about some linguistic effect.

The reliability and robustness are basic requirements for an automated machine translation system for processing UGC. An MT system for processing UGC should be 1) thoroughly customized for this specific type of content and 2) be able to translate large amounts of text in real time.

3. Aim and Objectives

The main challenge was to achieve high quality of translation. Since manual editing of each review was impossible, the website functionality required a high quality automated translation system that does not require human post-editing. About 80,000 reviews are added to the website weekly, so TripAdvisor required a technically accurate solution for processing large volumes of text. Another client's requirement was to translate the existing content (over 10 million reviews) within a short period. Due to the huge amount of data human post-editing of every single review was impossible. At the same time, UGC is a challenge for MT, since such texts are highly informal and typically contain a significant number of spelling, stylistic and punctuation errors that affect the MT results.

Another important client's requirement was an efficient quality estimation system integrated into the final MT solution. As TripAdvisor wanted to publish high-quality translations only, PROMT had to design an automated quality estimation system with a quality threshold.

The translation results had to contain clear and understandable content. Translation had to meet certain quality criteria, and as manual evaluation of the whole translation volume was impossible, the MT solution had to provide an automatic scoring mechanism for the evaluation of the translated texts.

The tight deadline for developing MT system was another crucial demand made by TripAdvisor.

Website developers wanted a cloud-based server MT solution, that's why we decided to develop a hybrid translation solution based on the PROMT DeepHybrid system (see [Molchanov, 2012]).

4. Statistical and Linguistic Analysis of the Data provided by TripAdvisor

4.1. Initial Data

TripAdvisor provided PROMT with the following data for engine customization:

- TripAdvisor English-Russian glossary (505 entries);
- English-Russian TripAdvisor TMs (~100,000 entries);
- English monolingual text corpus of hotel reviews (~1.2 billion words).

4.2. Domain-Specific Dictionaries

The TripAdvisor English-Russian glossary was converted into a dictionary of the PROMT internal format. We also extracted the most frequent terms and phrases from the English hotel review corpus. We analyzed the translations of these entries and made the necessary corrections and additions to the TripAdvisor dictionary and the baseline PROMT Travel dictionary.

Due to the large amount of misspellings and typos in the text of reviews we decided to create a dictionary with incorrect spelling of frequent English words, e.g.

(1) *couldnt, did'nt, experieince,*

so that the translation system could treat them as known words.

We also created the PROMT TripAdvisor Backgroud dictionary containing frequent travel-related phrases. The dictionaries were then icorporated into the translation system according to their priority: 1) TripAdvisor dictionary (highest priority); 2) TripAdvisor Backgroud dictionary; 3) Travel dictionary; 4) PROMT General dictionary (lowest priority). The priority works as follows: if the word or phrase is missing in the dictionary with the highest priority, the system tracks it in the dictionary with next priority etc.

4.3. Translation Memory

We made a thorough analysis of the English-Russian translation memory provided by TripAdvisor. We decided not to use it for three main reasons: 1) many segments were not domain-relevant; 2) many of them contained lots of different errors

(untranslated and incorrectly translated sentences, segments containing no alphabetic characters etc.); 3) many segments were of adequate quality but not informative for the baseline PROMT system, for example, named entities and geographic names:

(2) *Reno-PropertyOpen-NoDates Salute the white baroque towers of St. Fernando de Noronha and Atol das Rocas Reserves*

Due to the tight schedule we selected a random development set (approximately 10 percent) from the English hotel review corpus provided by TripAdvisor. We used this development set to build a list containing the most frequent in-domain sentences, e.g.

(3) *Highly recommended! The staff was very friendly and helpful.*

etc. These sentences (15K) were processed the following way: 1) the sentences were translated with the baseline PROMT system; 2) the translations were analyzed by our linguists. According to linguistic expertise only 8 % (1200 sentences) contained major syntactic and stylistic errors. These sentences were manually post-edited and integrated into the translation system as a translation memory.

4.4. Target Language Model

A target language model is normally built on the in-domain target texts. In our case, there was no in-domain text corpus in Russian, so we had to create it. We crawled and processed about 27,000 user reviews (80 million words) from different Russian websites dedicated to travelling. These texts were used to build the target language model. The model was integrated into the translation system.

A language model is a set of n-grams (word sequences of n-length) and their statistical characteristics. The rule-based system may have several translation options for some words and phrases. The language model is a component of the PROMT DeepHybrid system. It is used to score the translation candidates generated by the rule-based component and select the best one according to perplexity score. Perplexity (PPL) is inversely proportional to probability and is calculated for every translation candidate. The lower the PPL is, the better the translation candidate fits the language model.

We called the language model built on the Russian reviews corpus the BigTripAdvisor Language Model. It was integrated into the translation system for TripAdvisor.

4.5. Quality Estimation System

According to the client's requirements, our automated translation system had to be equipped with a quality estimation component. Quality estimation (QE) systems are used to estimate machine translation output quality at run-time. In our case, we had to select the high-quality translated reviews suitable for publishing on the website without human post-editing and reject the low quality ones.

First of all, we had to choose a confidence metric which would be the basic element of our QE system. Due to the tight schedule, we decided to create a simple metric based on PPL. Our experts performed the quality evaluation of 1000 sentences with different PPL scores. The results of this experiment showed that there is a sufficient correlation between the translation quality and the PPL scores (see Figure 1).

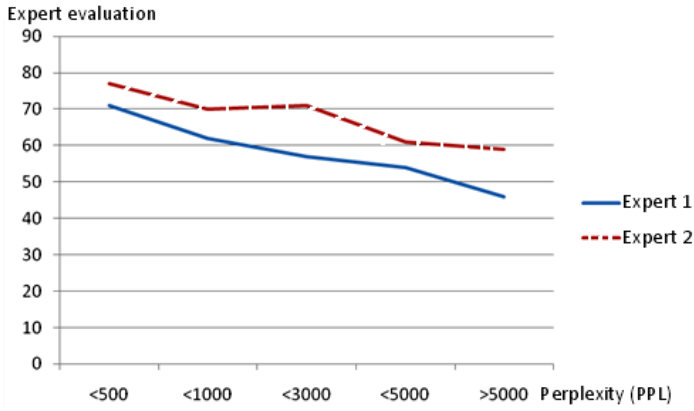

Fig. 1. Correlation between expert evaluation and PPL scores

QE systems normally operate on the sentence level. According to the client's request, our QE system had to estimate the entire text of the reviews. The average review length for the TripAdvisor website is approximately 100 words or three to five sentences. We decided to use the arithmetic mean of the PPL scores for separate sentences of reviews.

According to another request from TripAdvisor, the QE system had to be scaled from 1 to 5 with the accuracy of 0.1. Low-quality translations with PPL over 10,000 received the score equal to 1, high quality translations with PPL under 10 received the score equal to 5. The scaling formula is presented in Figure 2 below.

$$M = \begin{cases} 5, & PPL < 10 \\ \dfrac{4 \cdot (4 - \log_{10} PPL)}{3}, & 10 \leqslant PPL \leqslant 10^4 \\ 1, & PPL > 10^4 \end{cases}$$

Fig. 2. Scaling the PPL scores

We scored the translations of all reviews from the English monolingual corpus provided by TripAdvisor. The number of translations with scores 1 and 5 was less than 0.1 %. The distribution of the QE metric scores is presented in Figure 3.

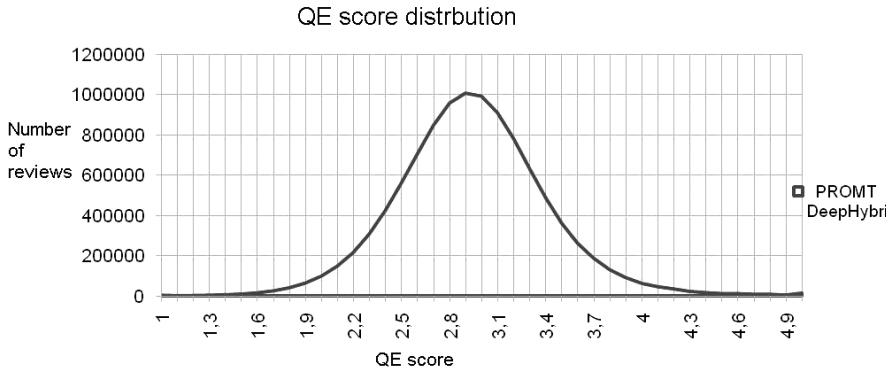

Fig. 3. Quality estimation score distribution

4.6. Deliverables

We developed a reliable, robust, scalable server-based translation solution. The solution was based on the PROMT DeepHybrid translation engine and included the following components:

- English-Russian Dictionaries: 1) PROMT TripAdvisor dictionary containing client-specific terms (approximately 5,600 entries); 2) PROMT TripAdvisor Background dictionary containing domain-relevant terminology (approximately 27,600 entries); PROMT TripAdvisor Geography background dictionary containing geographic names (approximately 48,200 entries).
- Target language model built on the text corpus of reviews in Russian.
- QE system.

5. Translation Quality Evaluation

Tripadvisor provided a parallel corpus (approximately 70K words) of the English reviews and their translations with human post-editing. We used this corpus to evaluate the translation quality. The English revews were translated with: 1) PROMT baseline system; 2) PROMT baseline system with the TripAdvisor dictionaries; 3) fully customized PROMT DeepHybrid system with all components. The BLEU scores are presented in Table 1.

Table 1. BLEU scores and the percentage of unknown words for various PROMT translation system configurations

System	BLEU score	percentage of unknown words
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%

Our experts performed lingustic analysis of the PROMT baseline system and the PROMT DeepHybrid system output. 3,291 sentences (78% of the test set) of the PROMT DeepHybrid system output contained changes compared to the PROMT baseline system output. Our experts compared 100 random RBMT and DeepHybrid translations in terms of improvements and degradations. The results showed that the DeepHybrid engine outperforms the RBMT engine according to human evaluation. The experts observed 49 improvements and 9 degradations for the DeepHybrid system output compared to the baseline system output. 42 translations were classified as equivalent.

Examples of translation quality improvements are presented in Table 2. The table also includes the translations of the Google online translation service.

Table 2. Examples of translation quality improvements

№	Source sentence	PROMT Baseline System	PROMT DeepHybrid system	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**.

6. Conclusions

We created an automated translation solution that fully answered the project objectives and the client's requirements. The entire process of system development and customization took about a month. The solution we created has the following features:

- Fast and efficient translation of large volumes of texts.
- High quality translation.
- Low costs for development and customization of the MT system (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.

We managed to show how an efficient MT solution for translating user-generated content can be developed and customized within a short period and with no parallel in-domain data.

References

1. *Banerjee P., Naskar S. K., Roturier J., Way A., Genabith J.* (2011), "Domain Adaptation in Statistical Machine Translation of User-Forum Data using Component Level Mixture Modelling", available at: http://mt-archive.info/MTS-2011-Banerjee.pdf
2. *Banerjee P., Naskar S. K., Roturier J., Way A., Genabith J.* (2012), "Domain Adaptation in SMT of User-Generated Forum Content Guided by OOV Word Reduction: Normalization and/or Supplementary Data?", available at: http://nclt.dcu.ie/mt/papers/Banerjee_EAMT_2012.pdf
3. *Flournoy R., Callison-Burch C.* (2000), "Reconciling User Expectations and Translation Technology to Create a Useful Real-world Application", available at: http://mt-archive.info/Aslib-2000-Flournoy.pdf
4. *Flournoy R., Rueppel J.* (2010), "One Technology: Many Solutions", available at: http://amta2010.amtaweb.org/AMTA/papers/4-05-FlournoyRueppel.pdf
5. *Jiang J., Way A., Haque R.* (2012), "Translating user-generated content in the social networking space", available at: http://amta2012.amtaweb.org/AMTA2012Files/papers/JiangWayHaque.pdf
6. *Molchanov A.* (2012), "PROMT DeepHybrid system for WMT12 shared translation task", available at: http://www.statmt.org/wmt12/pdf/WMT43.pdf