PROMT®

Automated Translation Solutions

Machine Translation of User Generated Content

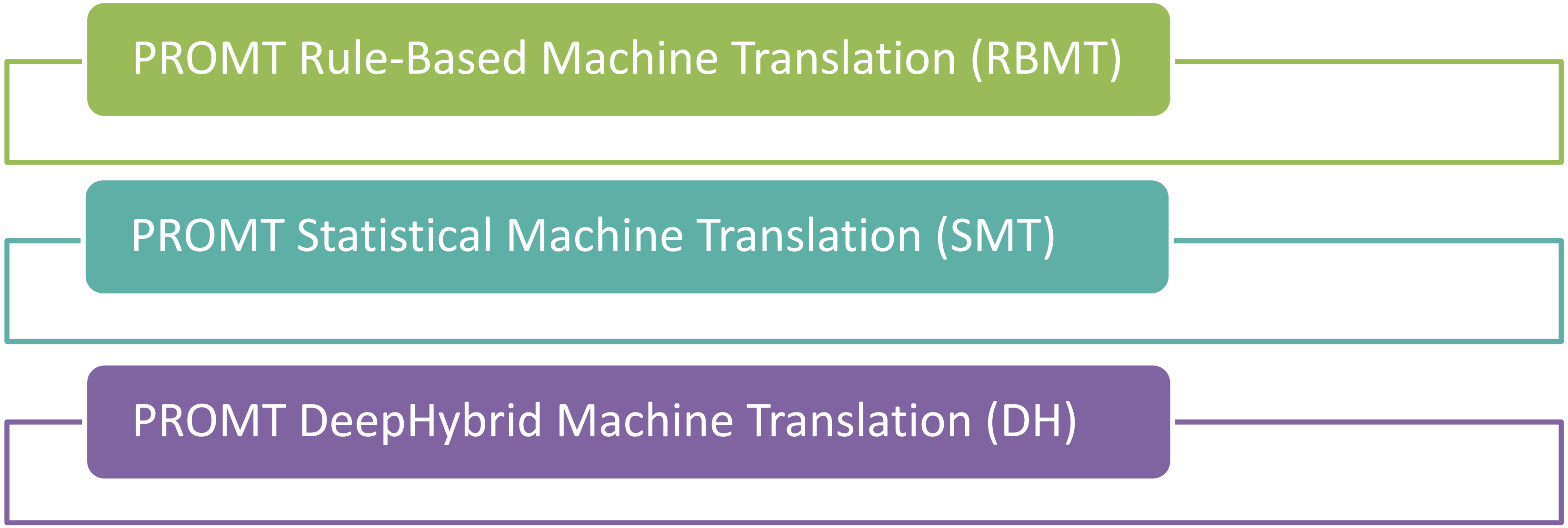
Julia Epiphantseva

Head of Business Development

TAUS

www.translationautomation.com

PROMT Technologies



- PROMT Rule-Based Machine Translation (RBMT)
- PROMT Statistical Machine Translation (SMT)
- PROMT DeepHybrid Machine Translation (DH)

Rule-Based Machine Translation

- Benefits:
  - more accurate syntax and morphology,
  - deterministic and predictable,
  - friendly for customization.
- Limitations:
  - language-dependent (algorithms depend on source/target languages),
  - high customization effort.
- Available languages in PROMT rule-based engines:

English, Russian, German, French, Spanish, Italian, Portuguese, Chinese (Simplified and Traditional), Ukrainian, Kazakh, Turkish, Bulgarian, Latvian and Polish.

- Available Products: Desktop and Server solution.

Statistical Machine Translation

- Benefits:
  - more fluent and “human-like” MT output,
  - language independent,
  - fast training.
- Limitations:
  - requires large and clean parallel corpora for training,
  - domain-specific (usually trained on/for specific texts),
  - requires powerful servers (slow).
- Available languages: language-independent.
- Available Products: Server-based solutions only.

PROMT DeepHybrid Machine Translation

- PROMT DeepHybrid takes the best from both approaches:
- Benefits:
  - more fluent and “human-like” MT output than pure RBMT,
  - engine training is fully automated
  - engine training is faster than pure RBMT,
  - more customizable and predictable then pure SMT.
- Limitations:
  - requires parallel corpora for training (but less than pure SMT),
  - domain-specific (usually trained on/for specific texts).
- Available languages in PROMT DeepHybrid: English, Russian, German, French, Spanish, Italian, Portuguese, Chinese (Simplified and Traditional), Ukrainian, Kazakh, Turkish, Bulgarian, Latvian and Polish.
- Available Products: Server-based solutions only.

User-generated content (UGC)

- produced by general public,
- available mostly on the Web via blogs and wikis,
- presented as daily news, encyclopedias, references, product or service reviews,
- important for social networking and eCommerce websites.

Could the output quality be improved through quick training?

UGC in linguistic aspect

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

Online services powered by PROMT

itranslate4.eu SpanishDict

voila.fr Translate.Ru tripadvisor

Subtitles as training data

- Advantages
  - available public (http://www.opensubtitles.org),
  - large or suitable amounts,
  - spoken, modern language.

- Disadvantages and risks
  - data quality,
  - compliance to the domain (traveling).

English-Spanish

Training data

- Size
  - ≈ 17 M parallel segments (sentences)
  - ≈ 110 M English words

- Data processing and filtering
  - normalizing punctuation, ligatures etc.
  - deleting duplicated, untranslated etc. segments

Test data

- Source
  - Traveler reviews and their Spanish human translations

- Size
  - 1 000 parallel segments
  - 15 500 English words

English-Russian

Training data

- Size
  - ≈ 3,4 M parallel segments (sentences)
  - ≈ 18 M English words
- Data processing and filtering
  - normalizing punctuation, ligatures etc.
  - deleting duplicated, untranslated etc. segments

Test data

- Source
  - Traveler reviews and their Russian human translations
- Size
  - 4 000 parallel segments
  - 67 000 English words

Evaluation results

Bleu scores

English-Spanish	English-Russian
34, 93 (RBMT) -> 38,58 (DH)	19,63 (RBMT)-> 19,06 (DH)

Expert evaluation for random 100 segments

English-Spanish	English-Russian
37% better	17% better
29% worse	29% worse
34% equal	54% equal

Comparison of training data ES/ER

- Unknown words in English parts
  - 0, 8% (ES)
  - 1% (ER)

Similar percentage of known words.

- Target vocabulary (Spanish and Russian sample subcorpora of comparable size)
  - 250 000 words (ES)
  - 500 000 words (ER)

Much more word forms in Russian corpus than in Spanish.
Poorer quality of Russian subcorpus than of Spanish (spelling errors).

- Expert evaluation of parallel subcorpora (500 random segments)
  - 9% - alignment mistakes and 9% - bad quality of “human” translation (ES)
  - 18% - alignment mistakes and 15% - bad quality of “human” translation (ER)

Poorer quality of English-Russian corpus than of English-Spanish (alignment/human translation).

Additional researches

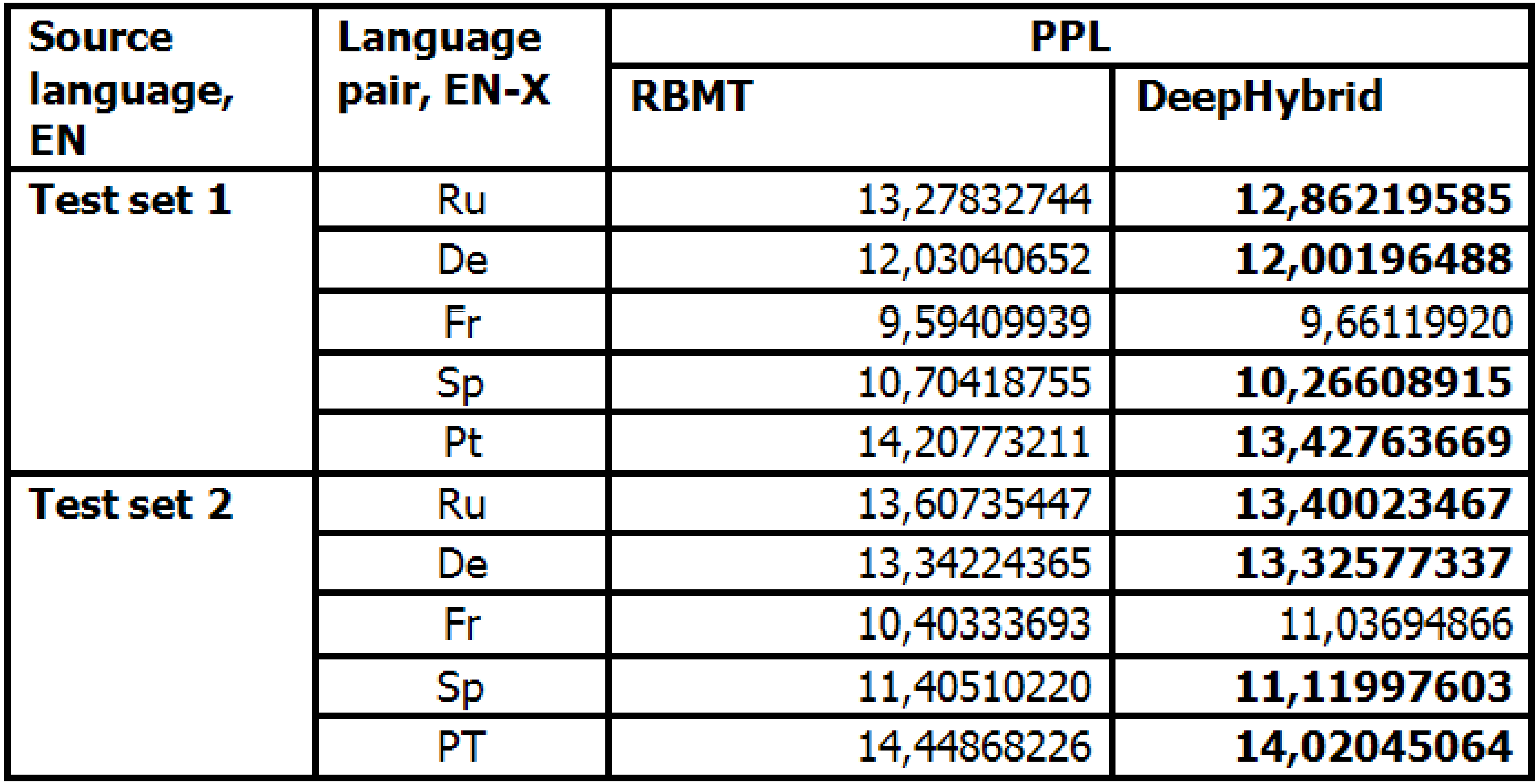
- More language pairs taken into consideration
  - English-French,
  - English-German,
  - English-Portuguese.
- Additional cleaning for training data
  - deletion of throw line marks at the beginning of segments,
  - validation of source-target sentences according to their length (1:1,5).
- Evaluation metrics
  - Expert evaluation
  - Language Model-based metric

Evaluation

Expert evaluation for random 100 segments

English-French	English-German
37% better	28% better
29% worse	20% worse
34% equal	52% equal

PPL Calculation



Source language, EN	**Language pair, EN-X**	**PPL**	
		RBMT	**DeepHybrid**
Test set 1	Ru	13,27832744	**12,86219585**
	De	12,03040652	**12,00196488**
	Fr	9,59409939	9,66119920
	Sp	10,70418755	**10,26608915**
	Pt	14,20773211	**13,42763669**
Test set 2	Ru	13,60735447	**13,40023467**
	De	13,34224365	**13,32577337**
	Fr	10,40333693	11,03694866
	Sp	11,40510220	**11,11997603**
	PT	14,44868226	**14,02045064**

Conclusions

- Translation quality
  - Improvement in translation output for Spanish/French/Portuguese
    - Romance languages are morphologically poorer than Russian,
    - no significant word-order differences between English and Romance languages,
    - Romance languages are more suitable for statistical approaches (SMT & Hybrid).
  - PPL rate reduction for all tested language pairs (except EF)
    - translation output became more "human-like" after training, but expert evaluation did not always confirm the real quality enhancement.
- Quality of training data
  - Open source data are always very noisy but substantial cleaning/filtering provides better results.
  - Subtitles are of especially bad quality,
  - More tools and approaches for data cleaning needed.

Thank you for your attention!

Julia Epiphantseva
Head of Business Development
Julia.Epiphantseva@promt.ru