达摩院 Alibaba DAMO Academy 以科技,创新世界

Quality Estimation Technology and its Applications in E-Commerce Machine Translation

Jiayi Wang
Machine Intelligence Technology Lab
DAMO Academy, Alibaba Group

Agenda

- Applications and Challenges of MT in E-Commerce Domain
- Multiple Strategies Improve E-Commerce Machine Translation
- Evaluation and Estimation of Translation Quality
- Human Evaluation of Translation Quality
- Automatic Translation Quality Estimation
- Applications of QE Technology in E-Commerce MT
- Conclusions

My Short Bio

- Currently working as a Senior Algorithm Engineer in the machine translation team of the Machine Intelligence Technology Lab at Damo Academy, Alibaba
- Responsible for developing real-time automatic quality inspection systems for online machine translation and speech recognition (ASR) engines.
- Received my Master Degree in the Department of Applied Mathematics and Statistics, Johns Hopkins University in 2015.
- Prior to Alibaba, I worked as a staff researcher in Social Science Research Institute (SSRI) at Duke University in the field of computational genetics from 2016 to 2017.

Applications of MT in E-Commerce

Website

Traffic Drainage	Insite Search	Preserve & Repurchase	Customer Conversion
Wide Coverage and Accuracy of Targeting	Accurate Translation of Query and Category	Trust & Service	Native Translation of Commodity Information

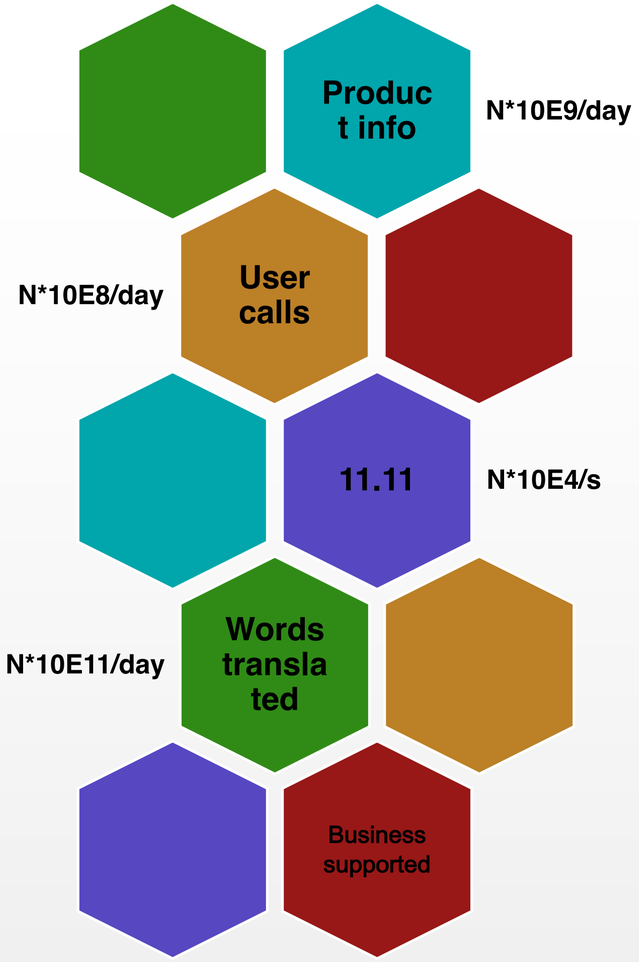
MT Service

Traffic Drainage	Insite Search	Preserve & Repurchase	Customer Conversion
1.Traffic Drainage Words 2.Landing Page 3.Creative Ideas	1.Language Identification 2.Spelling Correction 3.Query Translation 4.Category Translation	1.Information Rewriting 2.Title Translation 3.Description/Comment Translation	1.Overall Information Quality 2.Customer Service 3.Localization Service 4.Communication between Seller & Buyer

Business Index

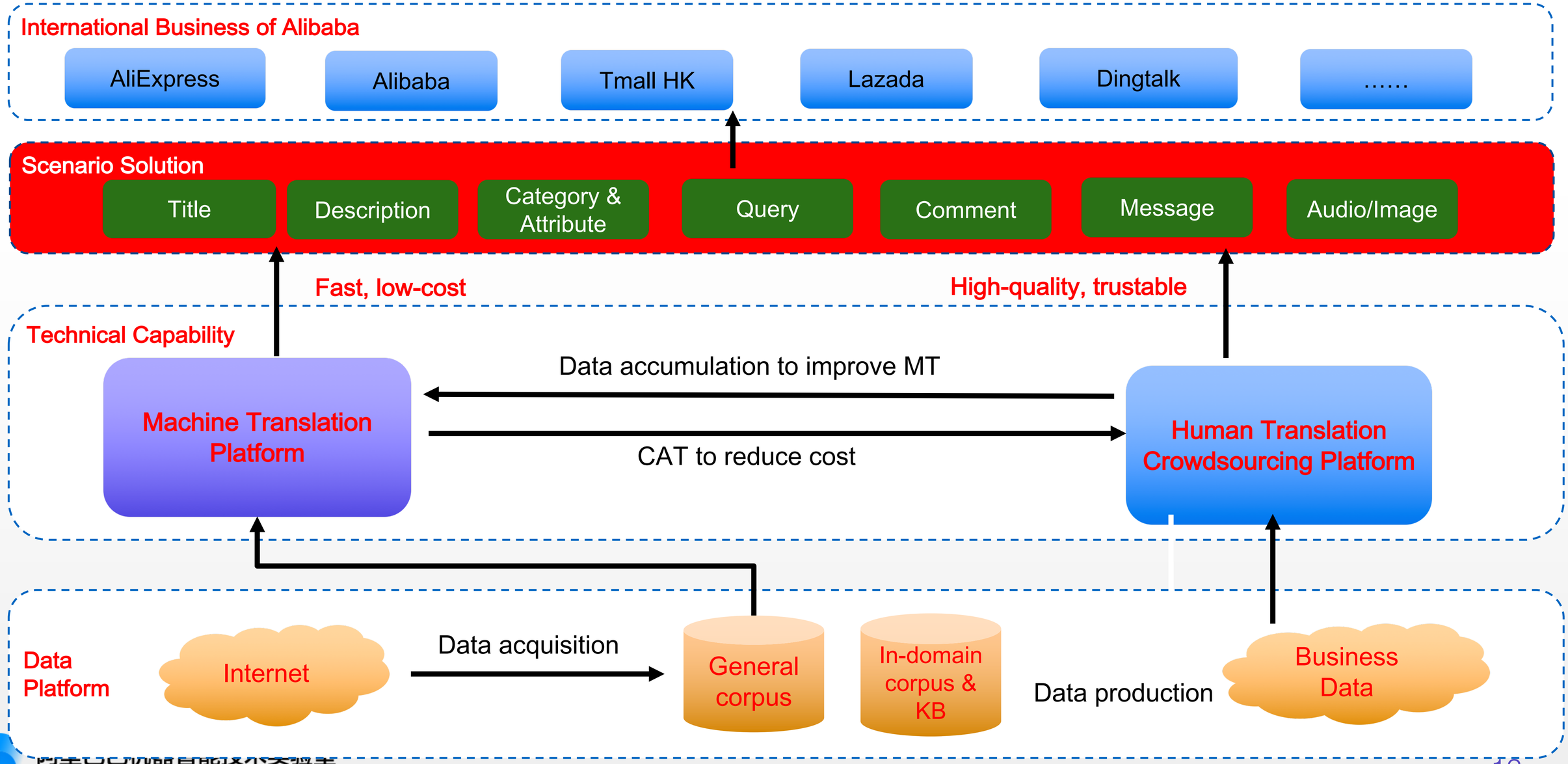
Traffic Drainage	Customer Conversion
UV ↑	GMV ↑
Unique Visitor	Gross Merchandise Volume

Alibaba MT's Capability



- Product information
  - X billions/day (offline)
- User calls
  - X billions/day
- 11.11 peak
  - X 10 thousands/second
- Translated words
  - X 100 billions/day
- Business supported
  - X 10 billions USD/year

Business Ecology of Alibaba Translation

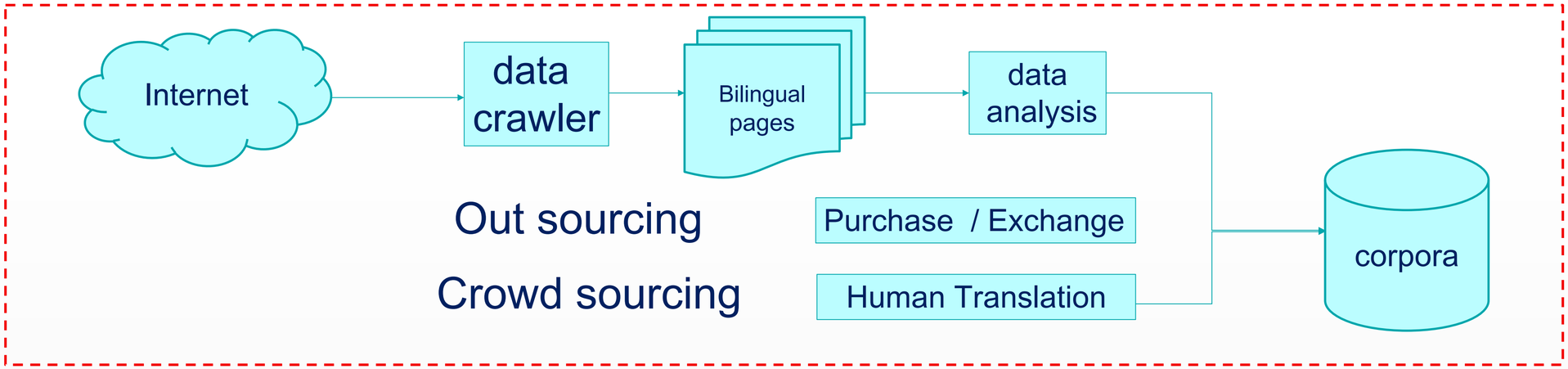


Challenges of MT in E-commerce

Translation Quality	•Readable output of target language •Accurate translation of key information •Flexible intervention mechanism
Speed	•Fast training on large-scale corpus •High concurrency •Low latency of inference
Service Quality	•High availability •Flexible and rich interface •Extendable to many language pairs •Efficient deployment and update

Data Strategy: Quantity

Parallel data



- Quantity
  - 20+ language pairs
  - Main language pairs, such as Chinese-English: N*10E8 (hundred millions)
  - Most of language pairs, such as Chinese-French: N*10E7 (ten millions)
  - Low-resource language pairs, such as Chinese-Vietnamese: N*10E6 (millions)

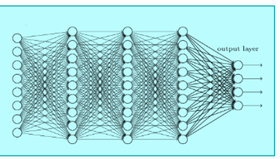
Data Strategy: Quality, Relevance

Quality Estimation

corpora

IBM model

RNN
force decoding



Data Classification

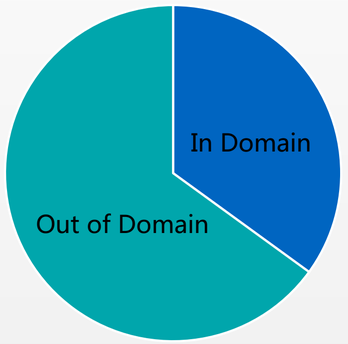
corpora

Data source information

Topic Model based data selection

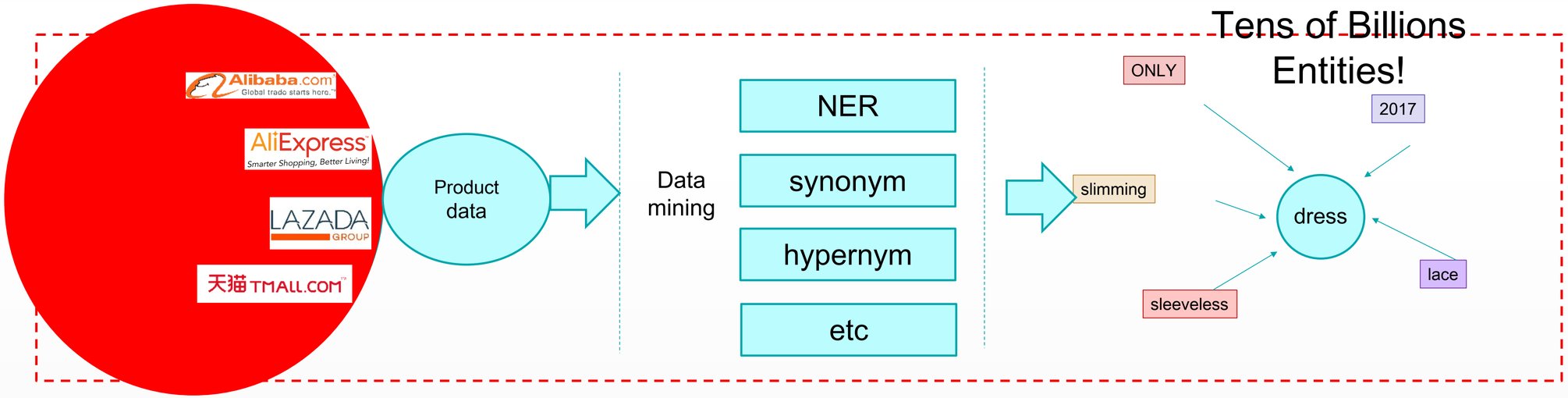
Language Model based data selection

CNN-based data classification



MT Knowledge Base

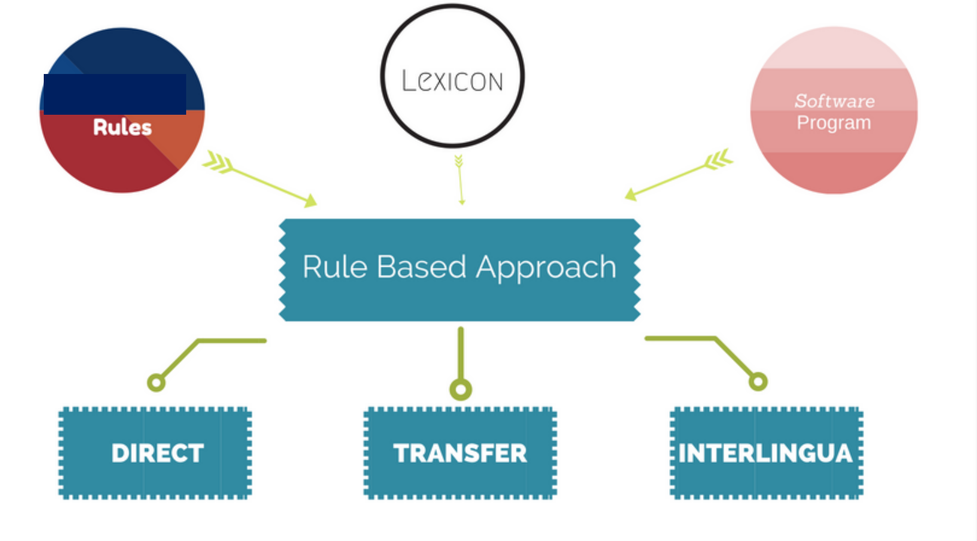
Ali Knowledge graph



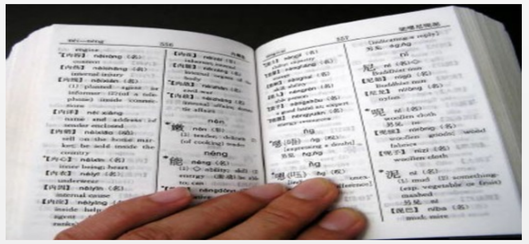
Multi-lingual knowledge base

MT Model Strategy: RBMT

- Rule Based Machine Translation
  - Direct matching
  - Map input to output with basic rules.
  - Rules are designed by human being.

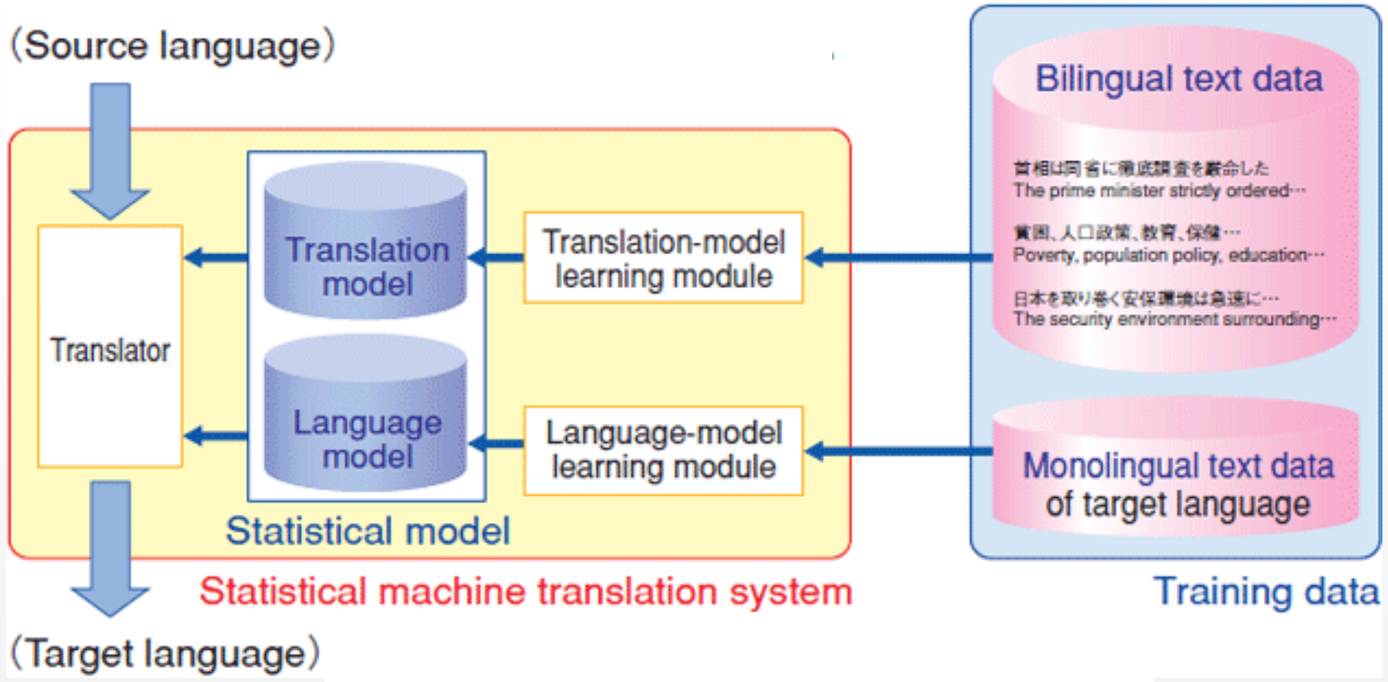


- Scenarios:
  - Numbers
  - Date
  - Address
  - Product info
  - NE



MT Model Strategy: SMT

- Statistical Machine Translation
- Scenarios:
  - Product title
  - Query

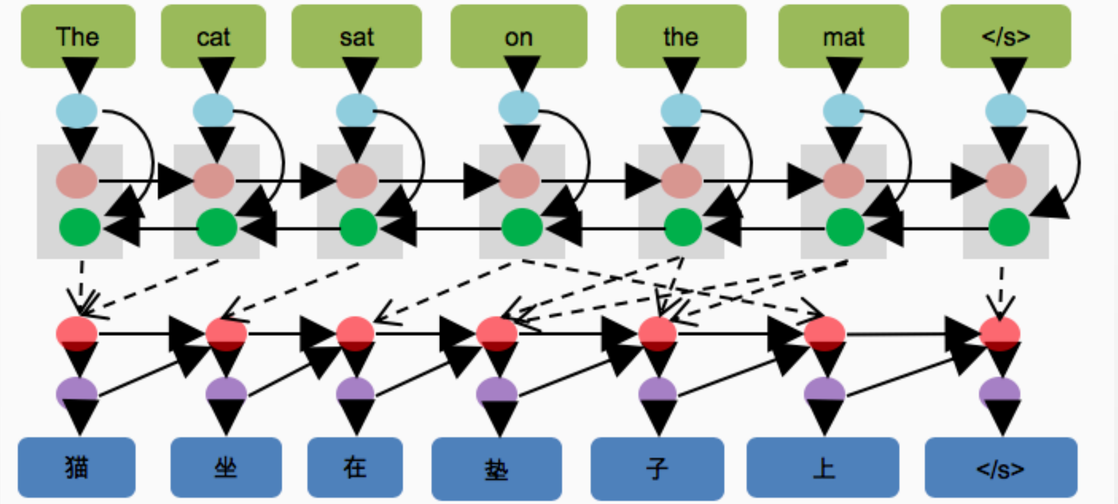


[Koehn, 2003]

MT Model Strategy: NMT

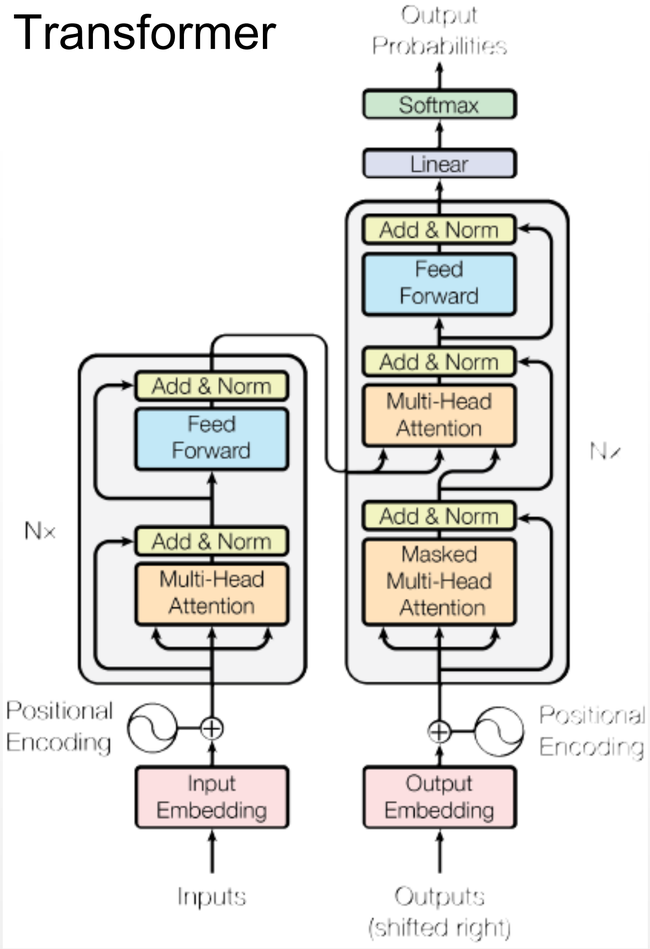
- Neural Machine Translation
- Scenarios:
  - Product description
  - Message
  - Offer
  - Comments

RNN-based
seq-seq model



[Bahdanau et al., 2015]

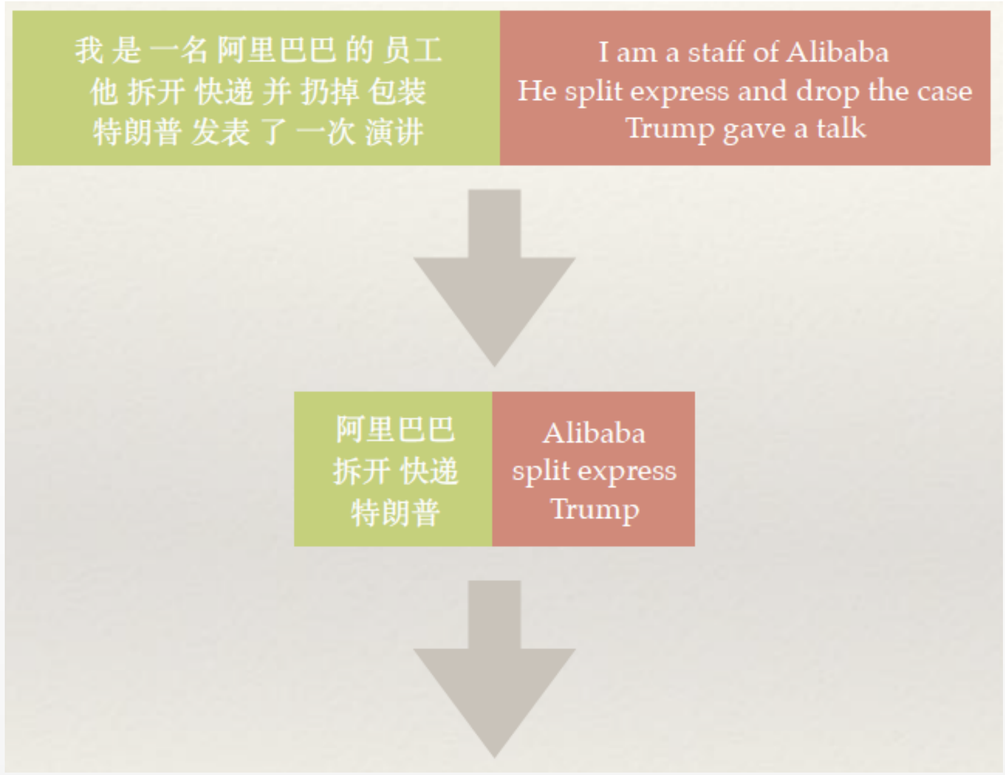
Transformer



[Vaswani et al., 2017]

MT Model Strategy: Constrained Translation

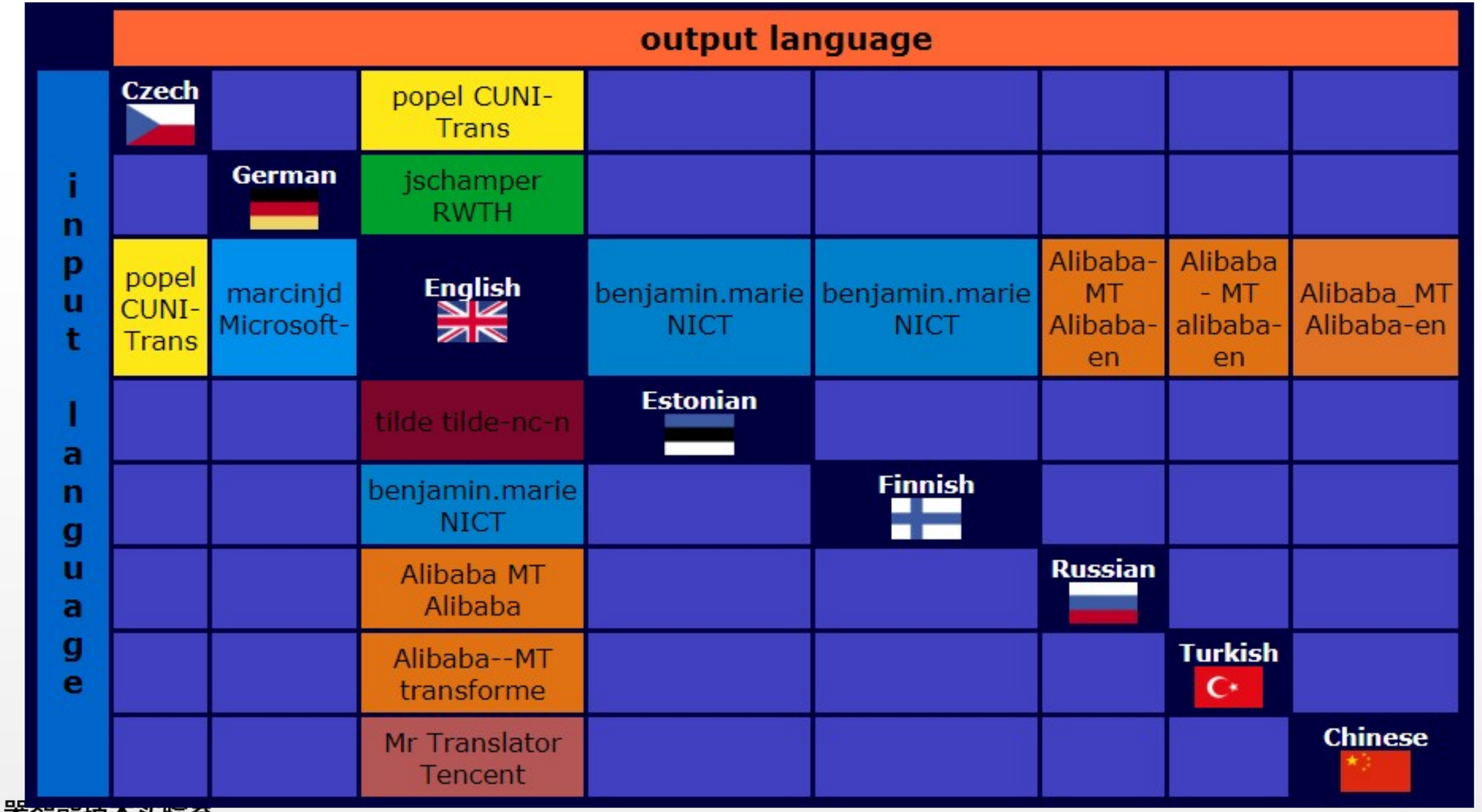
- A pre/post processing module for NMT
- Using attention information for alignment
- Trained using sentences with aligned tags
- Shared src and tgt tag embeddings.
  [Kuang et al., ACL 2018]



我 是 一名 <org> 阿里巴巴 </org> 的 员工
他 <VP> 拆开 快递 </VP> 并 扔掉 包装
<peroson> 特朗普 </person> 发表 了 一次 演讲

I am a staff of <org> Alibabab </org>
He <VP> split express </VP> and drop the case
<peroson> Trump </peroson> gave a talk

Alibaba @ WMT 2018 News Task



- Applications and Challenges of MT in E-Commerce Domain
- Multiple Strategies Improve E-Commerce Machine Translation
- Evaluation and Estimation of Translation Quality
- Human Evaluation of Translation Quality
- Automatic Translation Quality Estimation
- Applications of QE Technology in E-Commerce MT
- Conclusions

Translation Evaluation and Quality Estimation

- Translation Evaluation: measure the quality of the translations against golden references.

- Quality：
  - Fluency, adequacy?
  - System A vs System B?
  - Distance to a correct version?
  - How many major and minor errors ?

- Why do we need translation evaluation?
  - allow rapid comparisons between different systems.
  - enable the parameter tuning during system training.

Translation Evaluation and Quality Estimation

- Quality Estimation: estimate the quality of the translations without golden references.

- Quality:
  - Can we publish it as is?
  - Is it worth post-editing it?
  - Can a reader get the gist?
  - How much effort to fix it?

- Why do we need quality estimation?
  - quality control in translation industry
  - parallel data cleaning before system training

Automatic Evaluation Metrics

- No linguistic resource used
  - PER, WER, BLEU [Papineni et al., 2002], NIST [Doddington, 2002], RIBES [Isozaki et al., 2010], LRscore [Birch and Osborne, 2011], PORT [Chen et al., 2012], TER (translation edit rate) [Snover et al., 2006], CHRF
- Use limited linguistic information
  - Meteor [Banerjee and Lavie, 2005], TESLA [Liu et al., 2011], Meteor Universal [Denkowski and Lavie, 2014], TER-Plus [Snover et al., 2009], AMBER [Chen and Kuhn, 2011]
- Used higher level syntactic or semantic analysis
  - STM and DSTM [Liu and Gildea, 2005], DCU-LFG [He *et al.*, 2010], MEANT [Lo and Wu, 2011], Dreem [Chen et al., 2015]

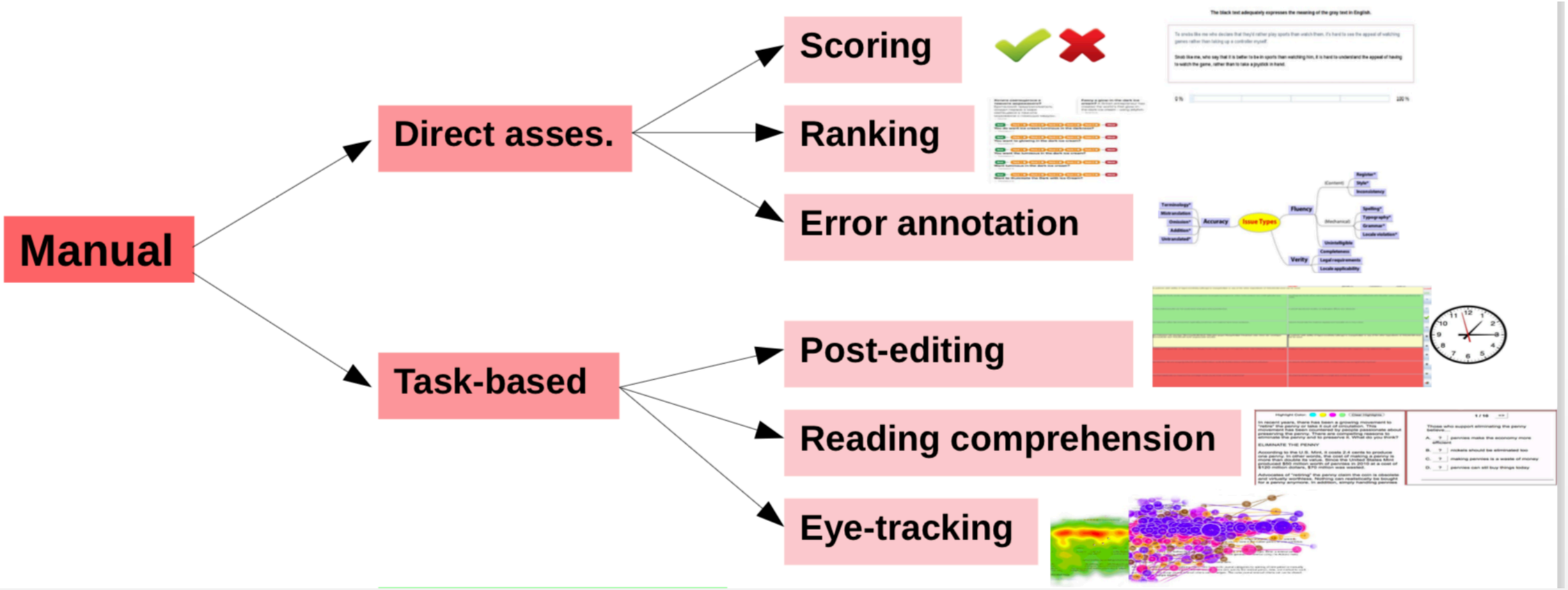
- BLEU and Meteor are two most widely used metrics.

Problems with Reference-Based Evaluation

- Requires human references
- Reference(s): only a subset of good translations
- Huge variation in reference translations.
- Metrics completely disregard source segment
- Cannot be applied for MT systems in use
- Increased score do not necessarily indicate improved translation quality

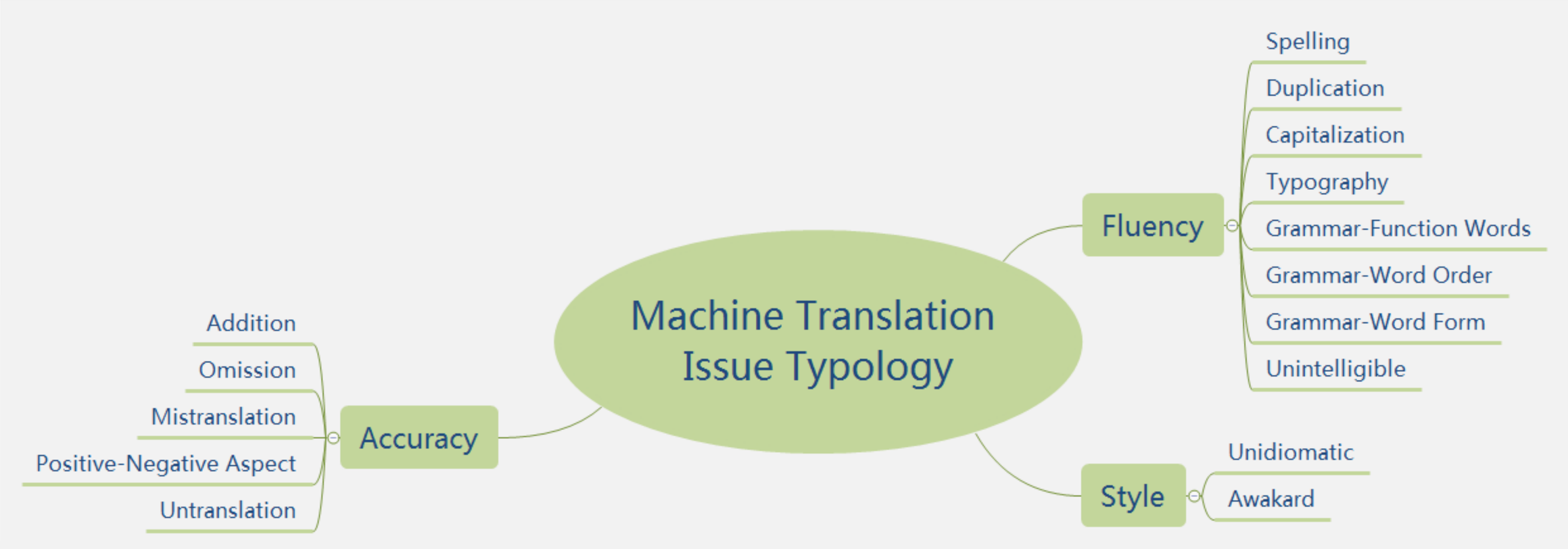
Credit: Lucia Specia

Human Evaluation Methods



Credit: Lucia Specia

Human Evaluation @ Alibaba



Human Evaluation @ Alibaba: LQI

Severities	Definitions
Minor	A Minor error is an error that may be noticed by a user, but will not confuse or mislead them. **For example:** Accuracy, font formatting, punctuation/spelling, grammar, syntax errors that produce a slight change in meaning, but not a loss of meaning
Major	Major issues are issues that impact usability or understandability of the content but which do not render it unusable. a) Will confuse or mislead the user and possibly bring him to write to User Support for help b) Will make the user suspect the quality of our products and start looking for other products as replacement. c) Incorrect links or email addresses d) Untranslated segments, blank translations e) Mistakes that are highly visible (home pages, promo etc.) or strike the eye at first reading f) Mistakes in our brand/product names
Critical	Critical issues are issues that render the content unfit for use. A Critical error is an error that can lead to: a) harm to the user or damage to a product b) financial consequences (loss of customers) c) a law-suit d) damage to Alibaba reputation, or e) Customer Compliant

Human Evaluation @ Alibaba

- Language Quality Index (LQI) Score

$$TQ = \left(1 - \frac{\#Minor * Severity_{minor} + \#Major * Severity_{major} + \#Critical * Severity_{critical}}{word_{count}}\right) * 100$$

$$TQ_{new} = \frac{TQ + C}{Scalar}$$

We used ratio 1:5:10. After adjustment, it implies $\boldsymbol{C = 100}$ *and* $\boldsymbol{Scalar = 2}$

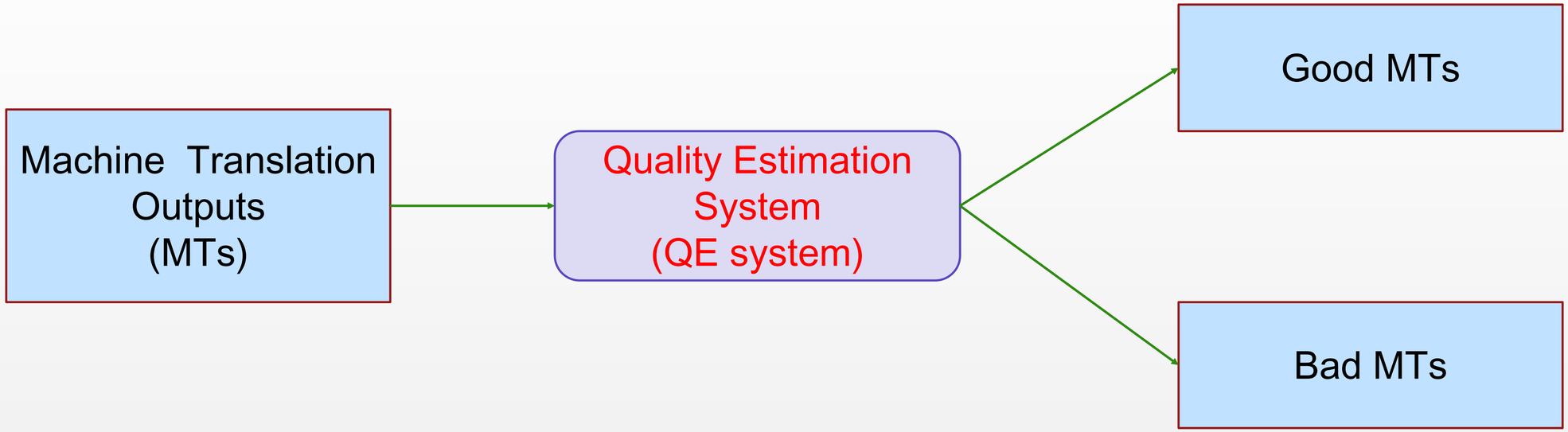
Disadvantages of Human Evaluation

- Time consuming, High cost
  - Only a small portion of translations can be evaluated
- Subjective
  - Low Inter- and Intra-annotator agreements

- Applications and Challenges of MT in E-Commerce Domain
- Multiple Strategies Improve E-Commerce Machine Translation
- Evaluation and Estimation of Translation Quality
- Human Evaluation of Translation Quality
- Automatic Translation Quality Estimation
- Applications of QE Technology in E-Commerce MT
- Conclusions

Automatic Quality Estimation

- Estimate the quality of translation at run-time;
- Estimate the quality of translation without any reference translation.



Sentence- & Word- Level Quality Estimation

- Sentence-Level
  - **Sentence Scoring** according to post-editing(PE) effort: percentage of edits need to be fixed (HTER)

- Word-Level
  - **Word Tagging** to predict OK/BAD tokens
    - number of tags = number of tokens
  - **Gap Tagging** to predict OK/BAD gaps (= predict missing tokens)
    - number of tags = number of tokens +1

For Example:

SRC: I have a red apple .
MT: 我 有 一个 粉 苹果 。
PE: 我 有 一个 红 苹果 。

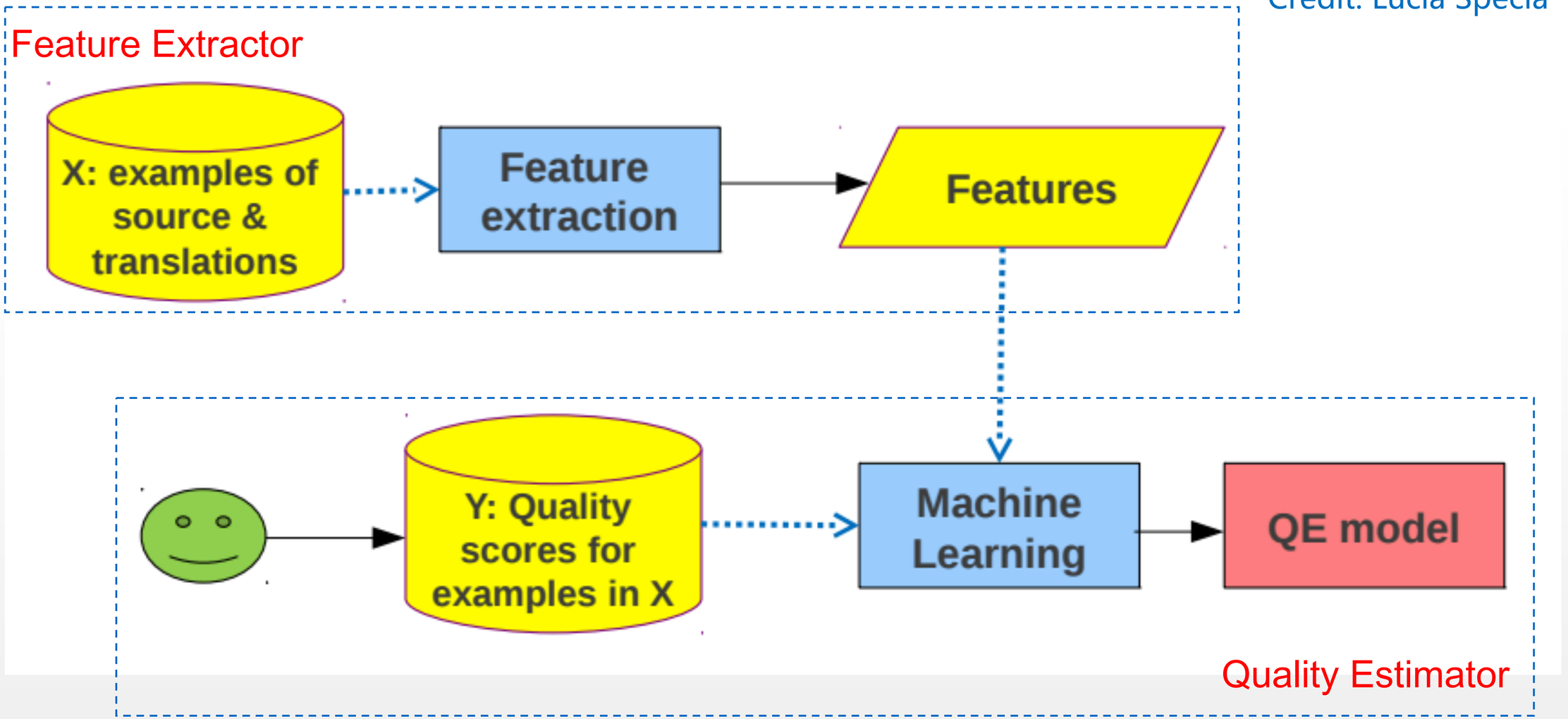
HTER: 1/6=0.167 (1 replacement)

Word Tags: OK OK OK BAD OK OK

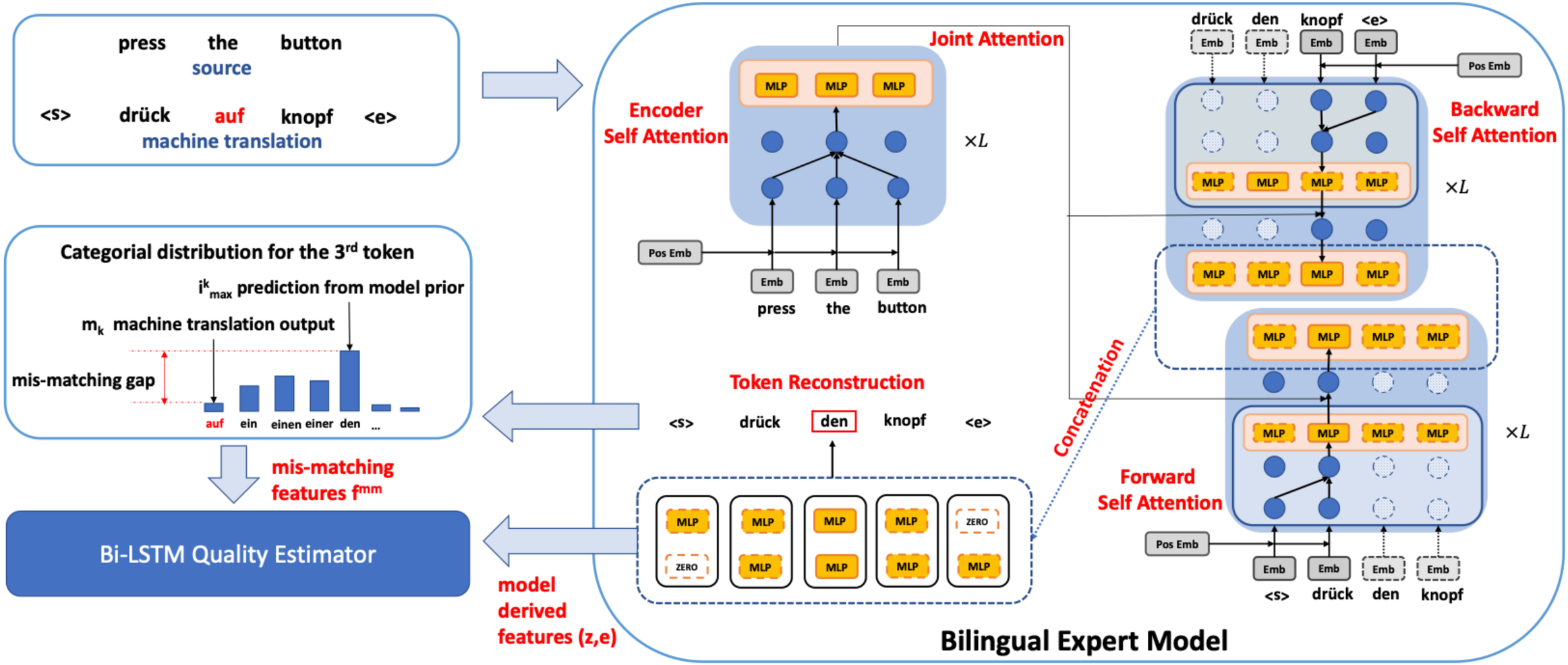
Gap Tags: OK OK OK OK OK OK OK

Quality Estimation System

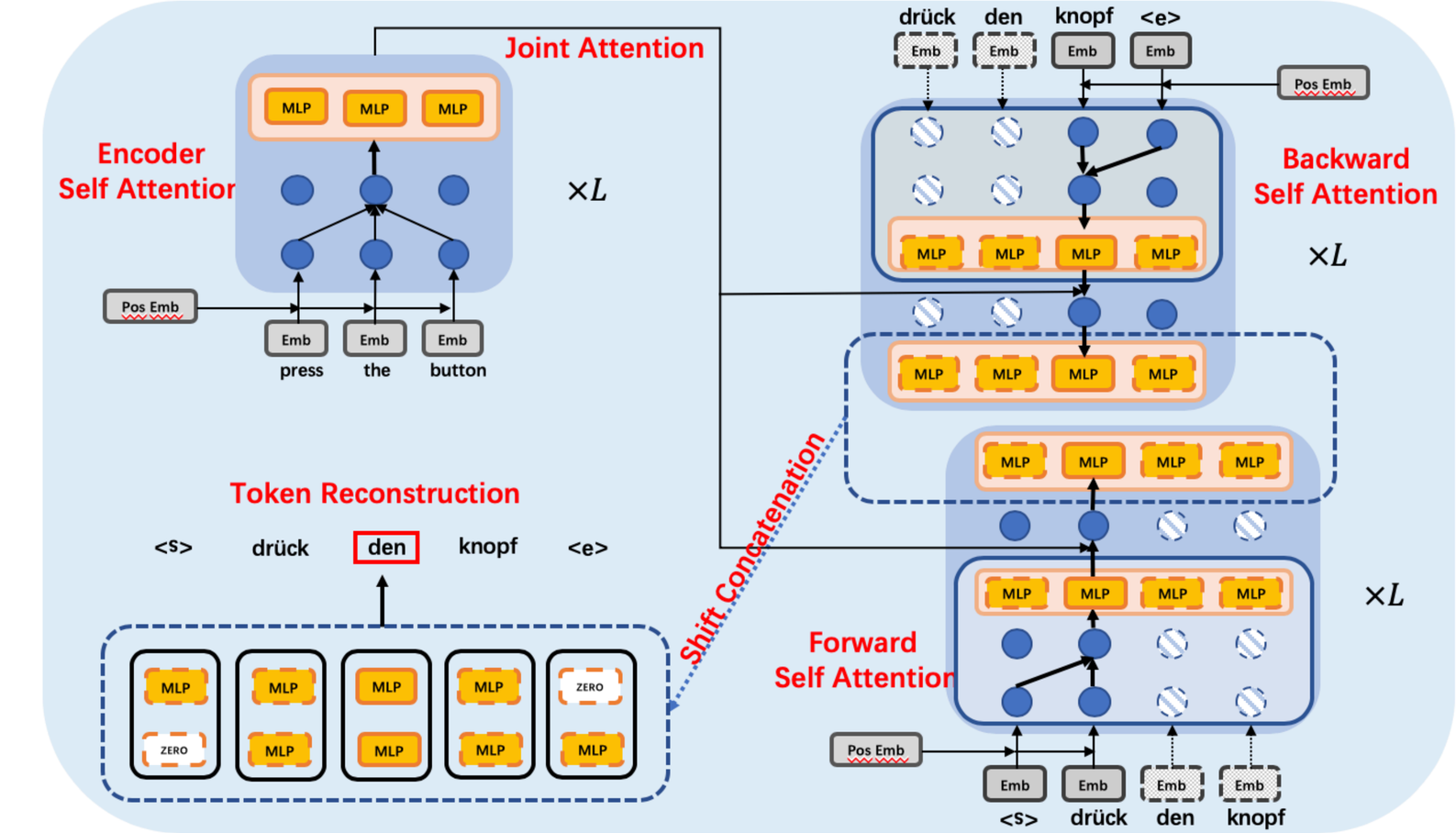
Credit: Lucia Specia



QE System @ Alibaba: QE Brain



Feature Extractor: Bi-directional Transformer



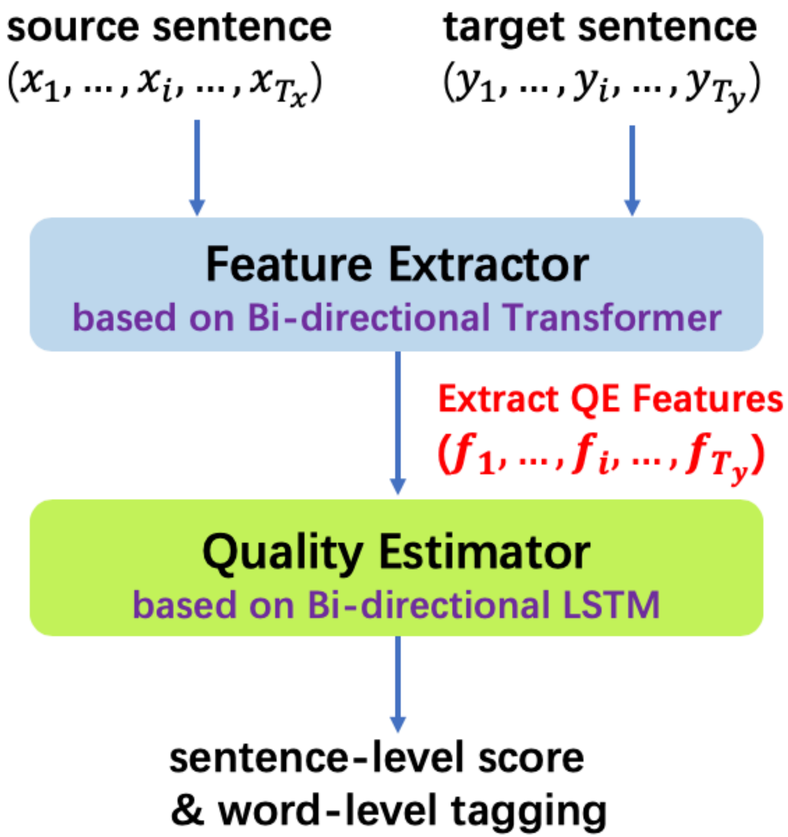
Three components:

(1)Self-attention encoder for the source

(2)Forward and backward self-attention encoders for the target sentence

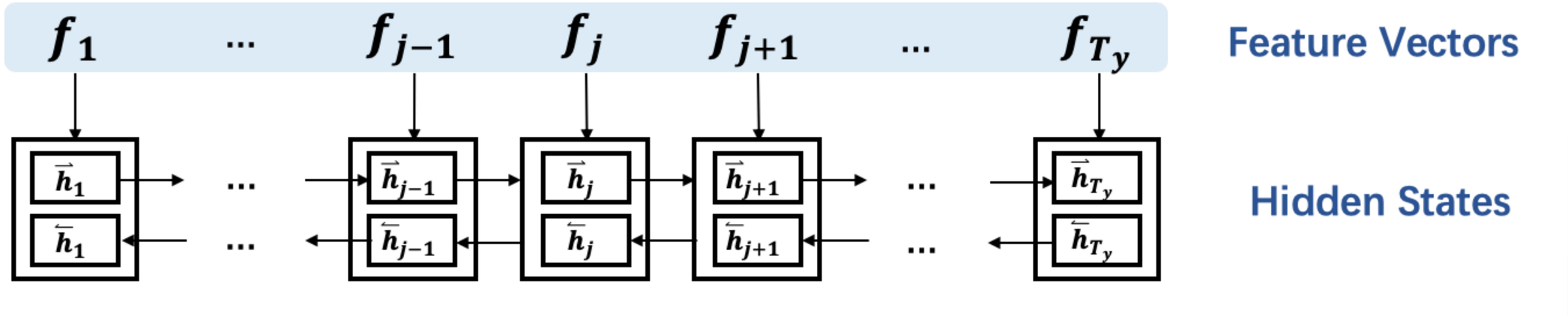
(3)The reconstruction for the target sentence

Features in QE Brain



- **Model Derived Features**
  - $\boldsymbol{z_k} = \boldsymbol{Concat}(\overrightarrow{\boldsymbol{z_k}}, \overleftarrow{\boldsymbol{z_k}})$, it contains the information from the source and the context around the k-th token. $\overrightarrow{\boldsymbol{z_k}}, \overleftarrow{\boldsymbol{z_k}}$ are sampled from $\boldsymbol{q}(\overrightarrow{\boldsymbol{z_k}}|\boldsymbol{s}, \boldsymbol{t_{<k}})$ and $\boldsymbol{q}(\overleftarrow{\boldsymbol{z_k}}|\boldsymbol{s}, \boldsymbol{t_{>k}})$
  - $\boldsymbol{e_k} = \boldsymbol{Concat}(\overrightarrow{\boldsymbol{e_{t_{k-1}}}}, \overleftarrow{\boldsymbol{e_{t_{k+1}}}})$
- **Mis-matching Features**
  - $\boldsymbol{p}(\boldsymbol{t_k}|\cdot)$ follows the categorical distribution with the number of classes equal to the vocabulary size.
  - $\boldsymbol{p}(\boldsymbol{t_k}|\cdot) \sim \boldsymbol{Categorical}(\boldsymbol{softmax}(\boldsymbol{l_k}))$ where $\boldsymbol{l_k}$ is the logits vector before applying the softmax operation
  - It achieves its maximum when $\boldsymbol{t_k}$ is ground true, so $\boldsymbol{p}(\boldsymbol{m_k}|\cdot) \leq \boldsymbol{p}(\boldsymbol{t_k}|\cdot)$ if $\boldsymbol{m_k} \neq \boldsymbol{t_k}$
  - Define 4-dimensional mis-matching features as $\boldsymbol{f_k^{mm}} = (\boldsymbol{l_{k,m_k}}, \boldsymbol{l_{k,i_{max}}}, \boldsymbol{l_{k,m_k}} - \boldsymbol{l_{k,i_{max}}}, \mathbb{I}_{\boldsymbol{m_k} \neq \boldsymbol{i_{max}}})$

Quality Estimator: Bi-directional LSTM



Concatenate the features along the depth direction to obtain a single one

- Sentence-level score can be formulated as a regression problem
- Word tagging prediction is a sequence labeling problem
- Gap tagging prediction is a sequence labeling problem

QE Model — Performance Boosting Strategy

Human-crafted Features
We introduce the human-crafted features as additional linear components for the predictive layer with a sigmoid activation function in the input of the Bi-LSTM quality predictive model

Fine-tune with Artificial QE Data
Round-trip translation in the APE task provides more supplementary training data, aiming to increase the diversity of erroneous translations during the training process so that it can reduce overfitting.

Greedy Ensemble Selection
The greedy ensemble selection algorithm, Focused Ensemble Selection (FES), helps to reduce the size of averaging ensembles but improve its efficiency and predictive performance.

Sentence-level	Test 2017 en-de		Test 2017 de-en	
	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ
QE single model	0.6837	0.7091	0.7099	0.6424
+HF	0.6842	0.7150	0.7085	0.6551
+FT	0.6957	0.7205	0.7128	0.6422
Ensembling	**0.7159**	**0.7402**	**0.7338**	**0.6700**

Word-level	F1-BAD	F1-OK	**F1-Multi**	F1-BAD	F1-OK	**F1-Multi**
Method	test 2017 en-de			test 2017 de-en		
QE Brain Base Single Model	0.6407	0.9045	0.5795	0.5750	0.9471	0.5446
+ FT	0.6410	0.9083	0.5826	0.5816	0.9470	0.5507
QE Brain Ensemble	0.6616	0.9128	**0.6039**	0.5924	0.9475	**0.5613**

QE Experimental Results @ WMT 2018

Sentence-level	Test 2018 en-de SMT		Test 2018 de-en SMT	
	Pearson's r	Spearman's ρ	Pearson's r	Spearman's ρ
Baseline	0.365	0.381	0.332	0.325
Competitor	0.700	0.724	**0.767**	0.726
Our System	**0.731**	**0.747**	0.763	**0.732**

Sentence-level	Test 2018 en-de NMT	
	Pearson's r	Spearman's ρ
Baseline	0.287	0.420
Competitor	**0.513**	**0.605**
Our System	0.501	**0.605**

Person's correlation:

- 0.0 No linear relationship
- 0.3 Weak linear relationship
- 0.5 Moderate linear relationship
- 0.7 Strong linear relationship
- 1.0 Perfect linear relationship

QE Experimental Results @ WMT 2018

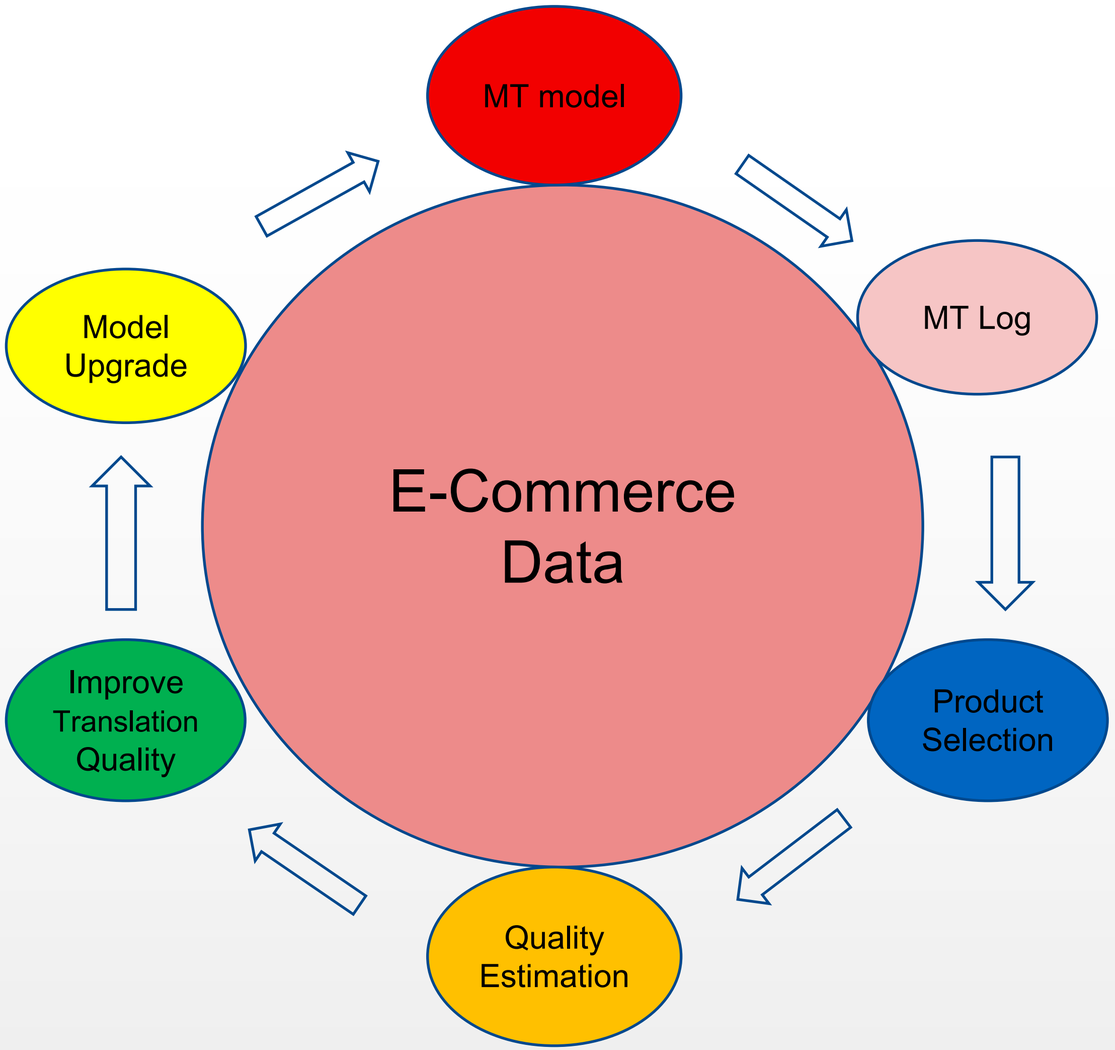
Word-level	Test 2018 en-de SMT	Test 2018 en-de NMT	Test 2018 de-en SMT
	F1-Multi	F1-Multi	F1-Multi
Baseline	0.363	0.181	0.437
Competitor	0.430	0.291	0.424
Our System	**0.607**	**0.435**	**0.593**

Alibaba @ WMT18 Quality Estimation Task

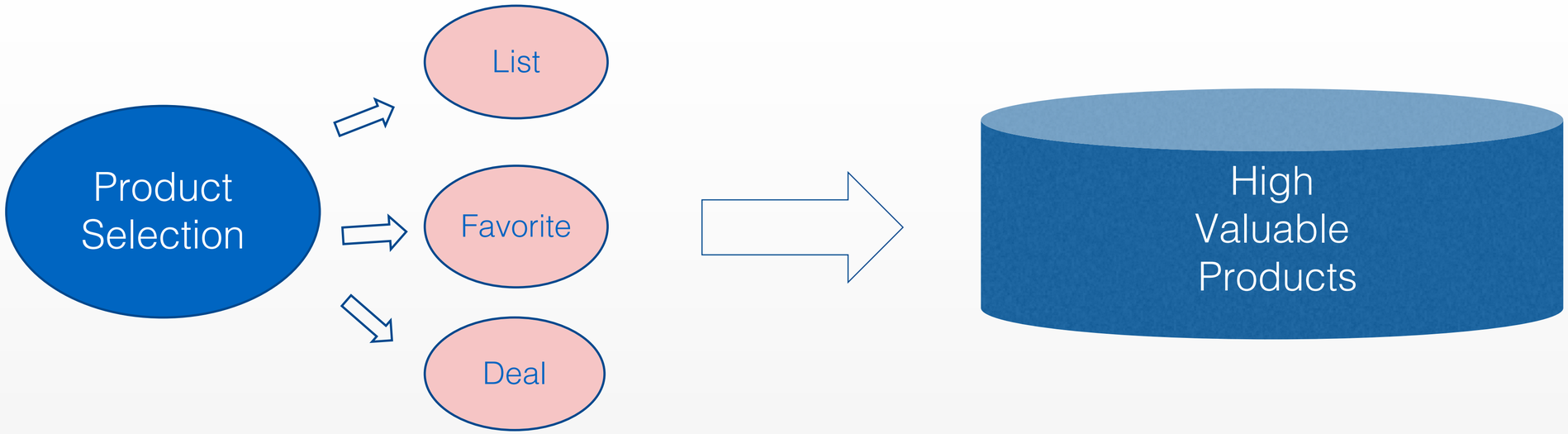
Alibaba@WMT18 QE	English-Germen SMT	English-German NMT	German-English SMT
Sentence-level	No. 1	No. 2	No. 1
Word-level	No. 1	No. 1	No. 1
Gap prediction	No. 1	/	/

- Applications and Challenges of MT in E-Commerce Domain
- Multiple Strategies Improve E-Commerce Machine Translation
- Evaluation and Estimation of Translation Quality
- Human Evaluation of Translation Quality
- Automatic Translation Quality Estimation
- Applications of QE Technology in E-Commerce MT
- Conclusions

QE in E-Commerce MT Loop

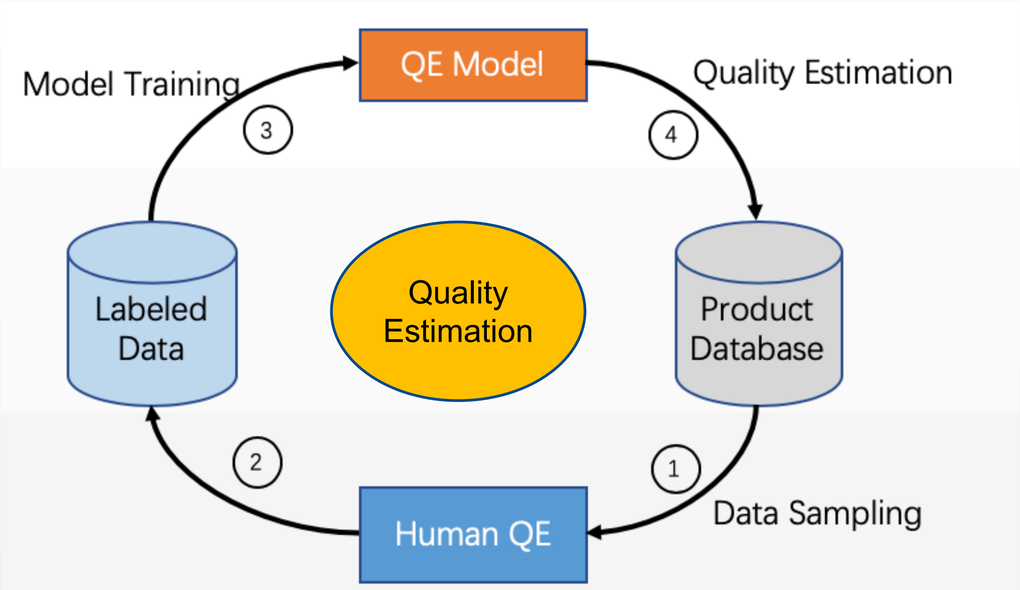


Product Selection Strategy

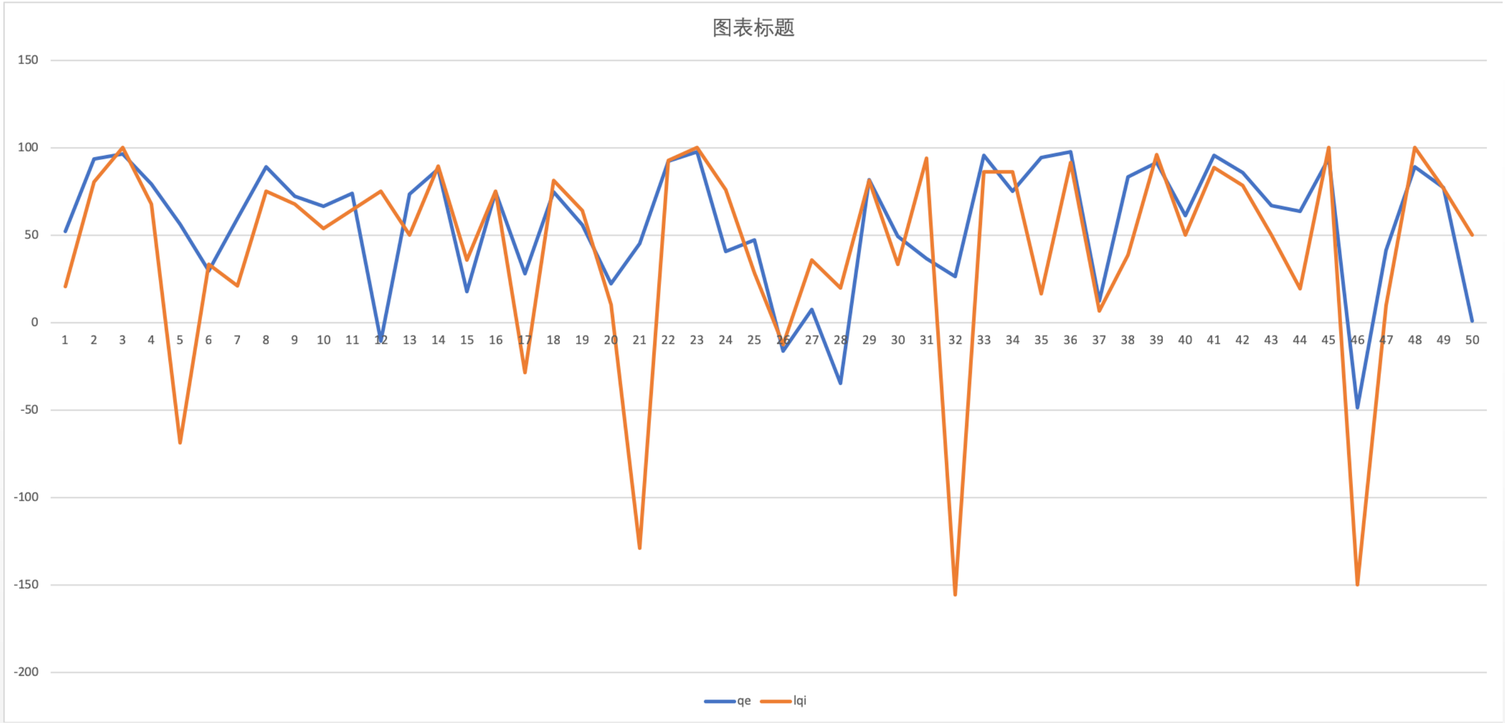


Quality Estimation

- Quality estimation according to different human translation quality measures
  - WMT: HTER score (previous slides)
  - E-Commerce: Linguist LQI score
  - E-Commerce: Scores based on users' behavior
  - E-Commerce: Crowdsourcing evaluation score

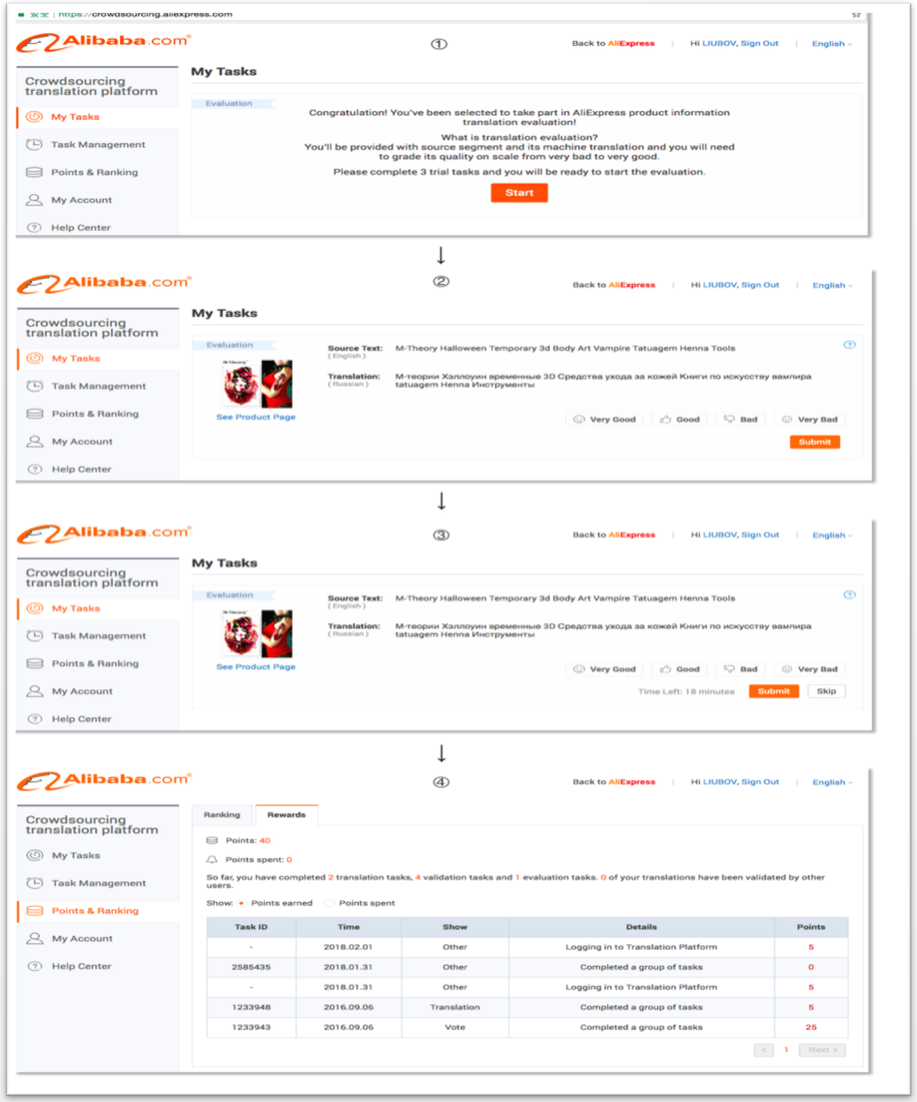


QE Based on LQI Score



QE Based on Crowdsourcing Evaluation

- Educated crowdsourcing evaluation

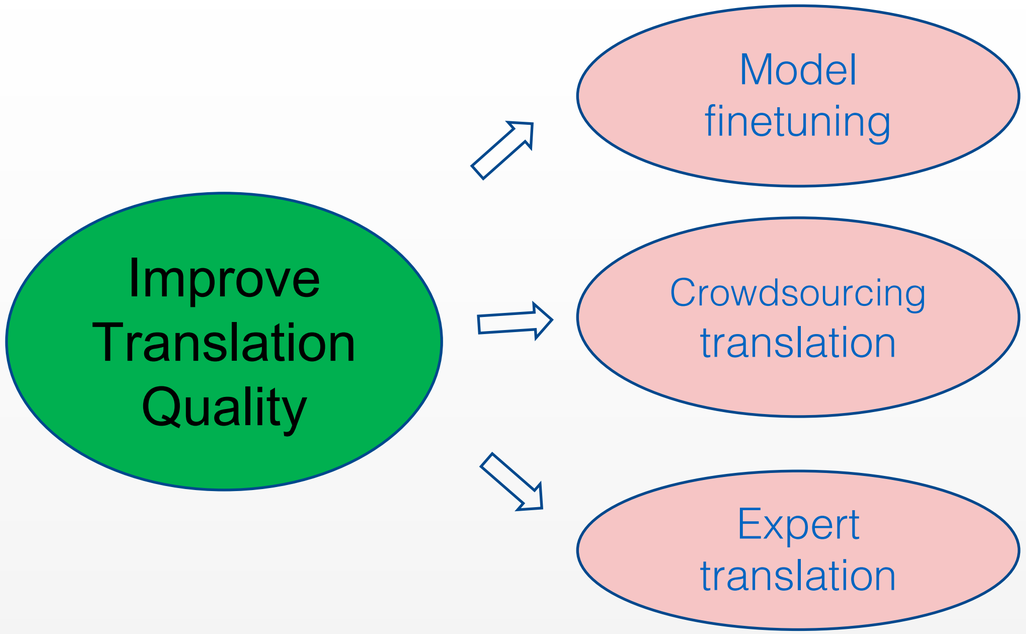


QE Based on Users' Behavior

- Pipeline of users' behavior

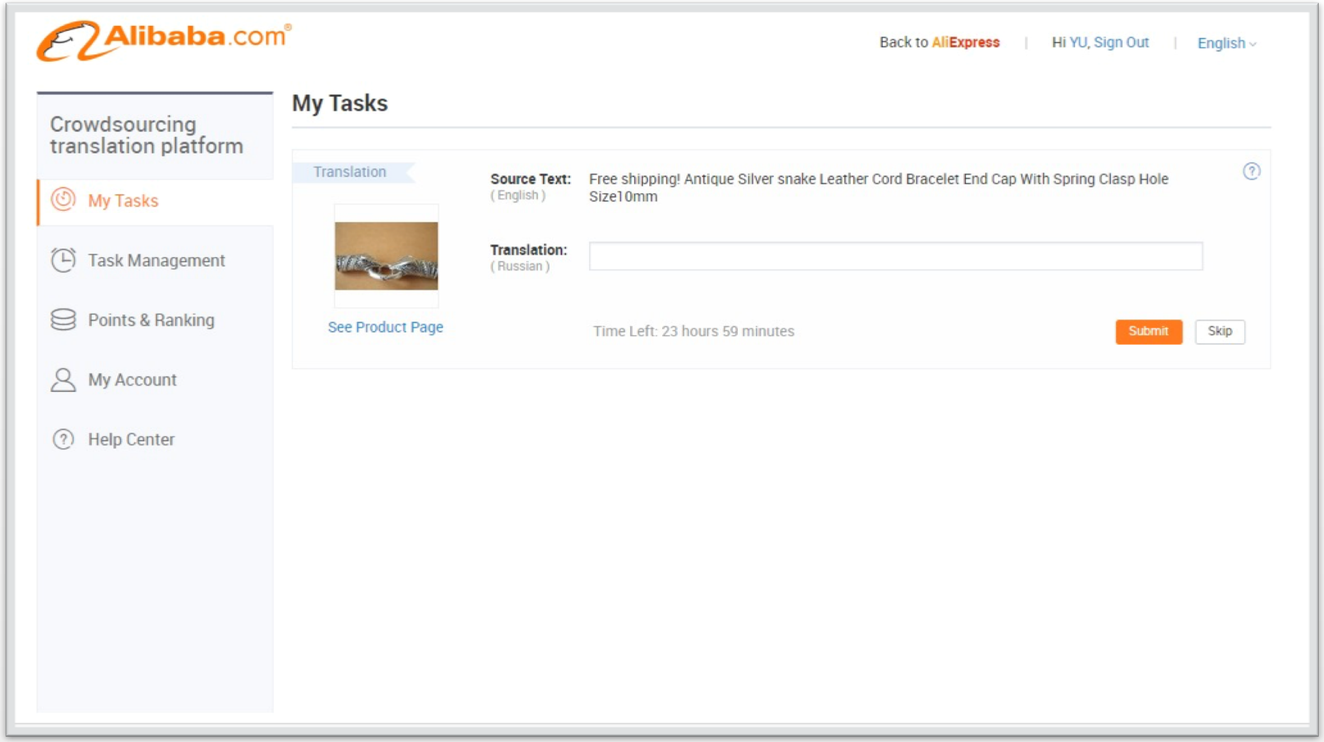
- Business index
  - UV
  - GMV

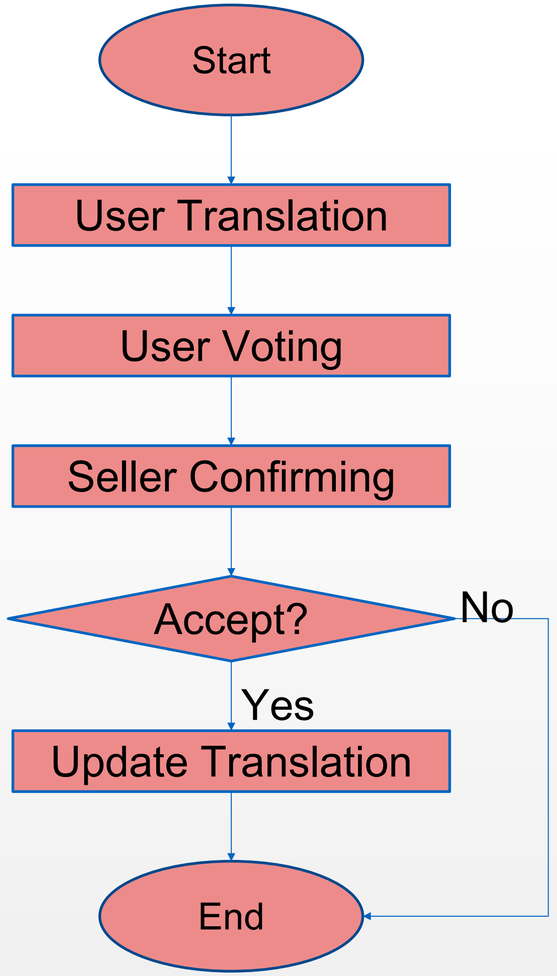
Improve Translation Quality



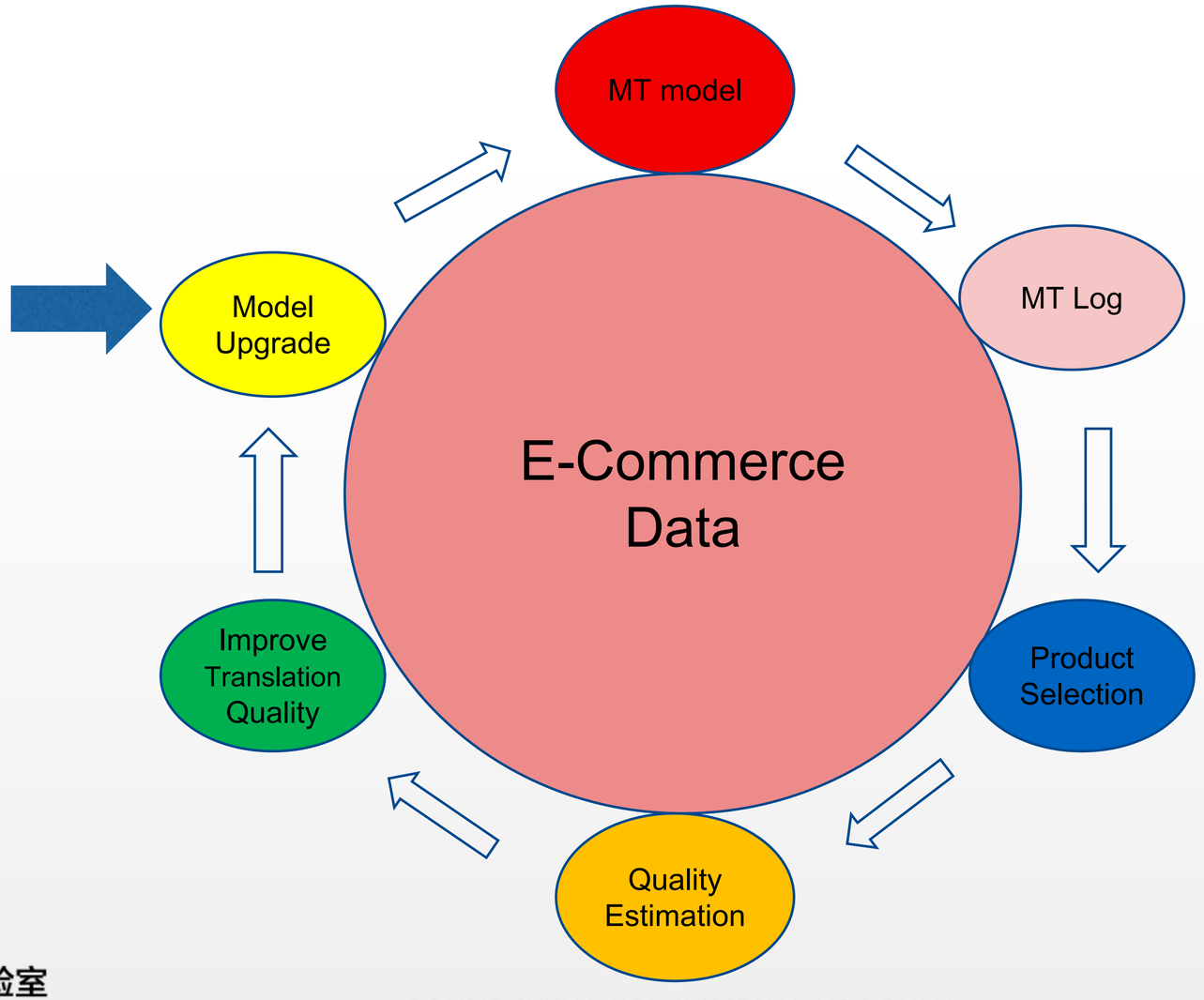
Crowdsourcing Translation

- Crowdsourcing Translation:
- http://crowdsourcing.aliexpress.com



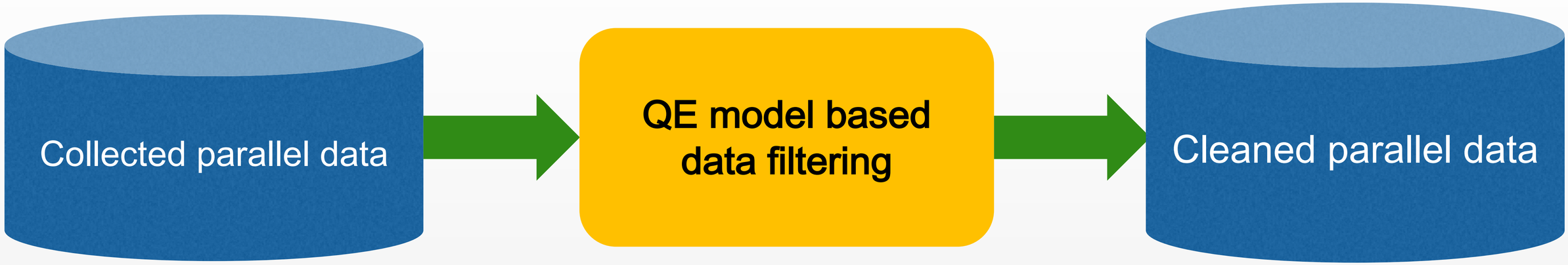


QE in E-Commerce MT Loop



Applications of QE Model: Training Data Filtering

- Training data quality is a key factor to NMT's performance.



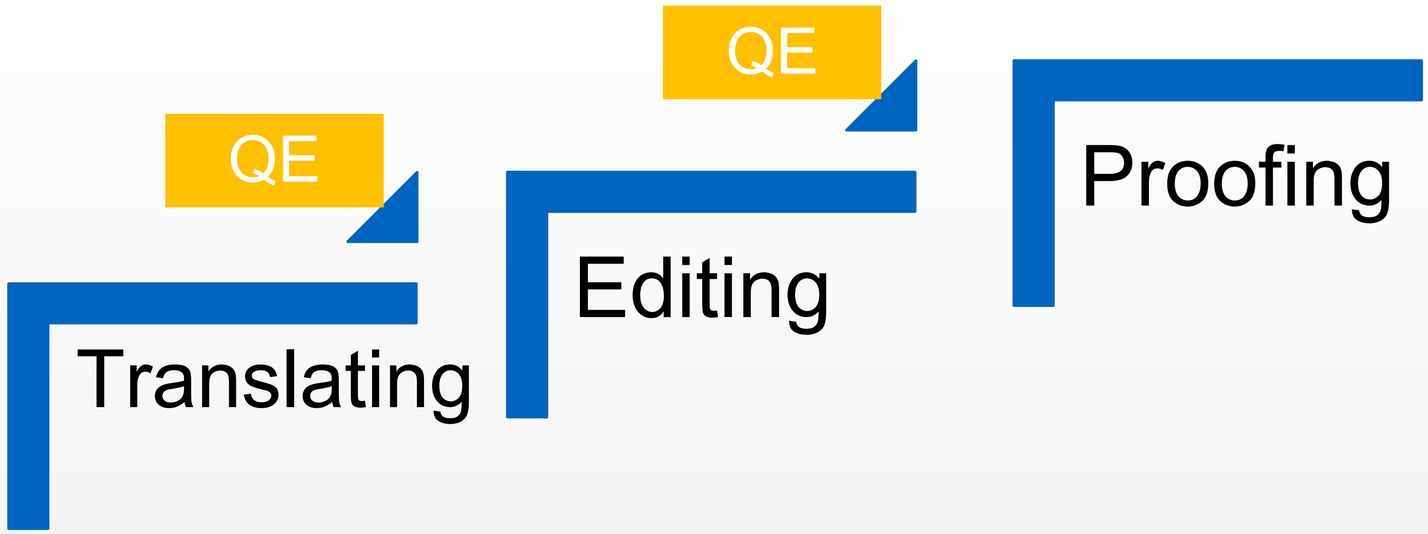
Applications of QE Model: Automatic Post-Editing

- Unify the Quality Estimation and Automatic Post-editing
- We can predict both words & gaps when condition on the entire source and the context in the target.

MT	wählen sie im bedienfeld ” profile ” des dialogfelds ” preflight ” auf die schaltfläche ” längsschnitte auswählen . ”
APE	klicken sie im bedienfeld ” profile ” des dialogfelds ” preflight ” auf die schaltfläche ” profile auswählen . ”
PE	klicken sie im bedienfeld ” profile ” des dialogfelds ” preflight ” auf die schaltfläche ” profile auswählen . ”
MT	sie müssen nicht auf den ersten punkt , um das polygon zu schließen .
APE	sie müssen nicht auf den ersten punkt klicken , um das polygon zu schließen .
PE	sie müssen nicht auf den ersten punkt klicken , um das polygon zu schließen .
MT	sie können bis zu vier zeichen .
APE	sie können bis zu vier zeichen eingeben .
PE	sie können bis zu vier zeichen eingeben .
MT	die standardmaßeinheit in illustrator beträgt punkte (ein punkt entspricht .3528 millimeter) .
APE	die standardmaßeinheit in illustrator ist punkt (ein punkt entspricht .3528 millimeter) .
PE	die standardmaßeinheit in illustrator ist punkt (ein punkt entspricht .3528 millimetern) .

Applications of QE Model: Human Translation Quality Control

- Integrating to human translation quality control process



Conclusions

- Machine translation is not perfect, but it can bring in real values in some scenarios, such as cross-border e-commerce, etc.;
- RBMT, SMT, NMT and constrained translation all have advantages in different scenarios of e-commerce translation;
- Alibaba has state-of-the-art machine translation technology.
- Bidirectional transformer is a strong representation encoder. QE system based on bidirectional transformer achieved the best results.
- Quality estimation is a crucial component for industrial machine translation, and QE based MT improvement loop helps e-commerce translation.
- We are open for collaboration!