Detecting Failures of Neural Machine Translation In the Absence of Reference Translations

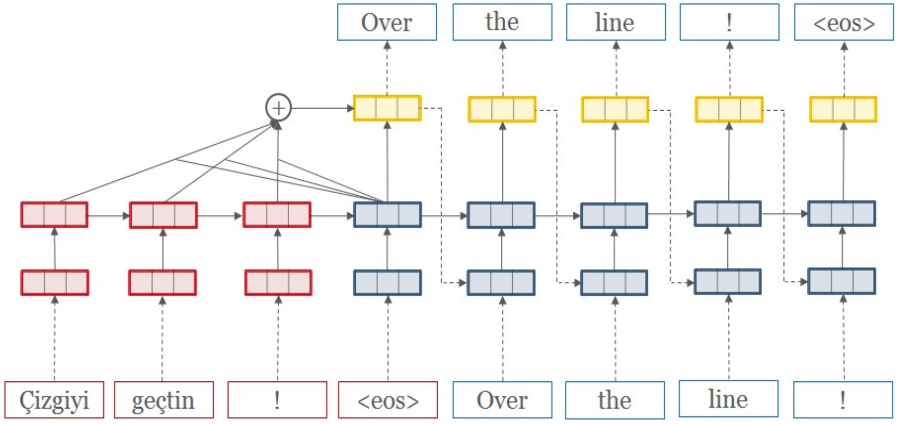
Wenyu Wang[UI] Wujie Zheng[TC] Dian Liu[TC] Changrong Zhang[TC]

Qinsong Zeng[TC] Yuetang Deng[TC] Wei Yang[UT] Pinjia He[EZ] Tao Xie[UI]

ILLINOIS Tencent腾讯 UTD ETH zürich

Neural Machine Translation (NMT)

- $argmax_{d_1,d_2,\ldots} P(d_1, d_2, \ldots | s_1, s_2, \ldots)$
- Statistical models -> neural networks
- Extensively researched & widely adopted
  - Satisfactory performance
  - Simpler architectures



Source: http://opennmt.net/

NMT Systems Can Be Error-prone

- Translation failures instead of software failures
  - Incorrect word/phrase translations
  - Incorrect semantics
  - … and many more
- Consequences are generally undesirable
  - Unsatisfactory user experience
  - Severe reputational and/or financial loss
- Still widely existing…

Source: https://www.k-international.com/blog/translation-fails-2018/

10 Hilarious Translation Fails From 2018

June 14, 2018 / 1 Comment / in Language Blog / by Richard Brooks

Source: https://www.rws.com/insights/rws-moravia-blog/ eight-of-the-most-bizarre-translation-fails-of-2018/

8 NOV 2018 | RWS MORAVIA BLOG |

TOPICS: JUST FOR FUN / MACHINE TRANSLATION (MT) / TRANSLATION /

Eight of the Most Bizarre Machine Translation Fails of 2018

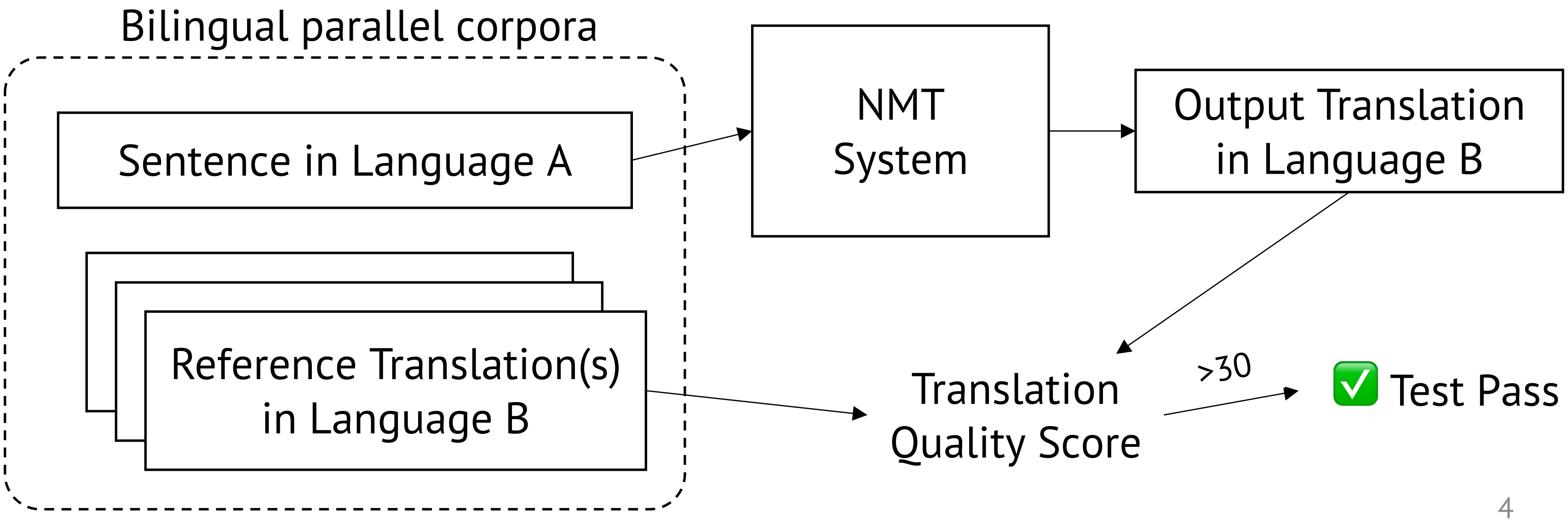
Source: https://www.searchenginepeople.com/blog/10-google-translate-fails.html

searchenginepeople

10 Inexplicable Google Translate Fails

NMT Quality Assurance: Common Practice

- *Reference-based* black-box system testing
  - Performed during in-house development
  - Evaluate on human-made bilingual parallel corpora
  - Calculate and observe translation quality indicators (e.g., BLEU scores)


Bilingual parallel corpora
Sentence in Language A
Reference Translation(s) in Language B
NMT System
Output Translation in Language B
Translation Quality Score
>30
✅ Test Pass

NMT Quality Assurance: What About Being Reference-free?

- Desirable benefits in industrial settings
  - Helping with translation quality improvement on more data
  - Enabling *in-vivo* testing and continuous monitoring in the production environment
  - Handling translation failures gracefully
- Existing approaches do not fulfill such demand

➢We aim for a practical and scalable solution to this challenge for our product

Reference-free Translation Failure Detection: Our Approach

- Focus on the 1-to-1 constituent mapping property of translation
  - Can be checked systematically
- Leverage both original texts and translated texts
  - As opposed to reference-based approaches
- Hybrid property violation detection strategy
  - Both statistical and systematic analysis

Constituent Mapping Property

- Constituents (e.g., words/phrases) are generally 1-to-1 mapped
    - Between the original text and the translation
- Any violation of this property in the translation indicates potential translation failures
- Two types of violations: under- and over-translation
    - Many translation failures can be reflected through these two types of violations

Under- and Over-translation

- Under-translation: words/phrases from the original text are missing in the translation

Chinese (original)	English (translated)	English (reference)
三姑给你的红包 给你妈妈了	Third Aunt gave you a red envelope.	Third Aunt gave your red envelope to your *mother*.

Example of under-translation

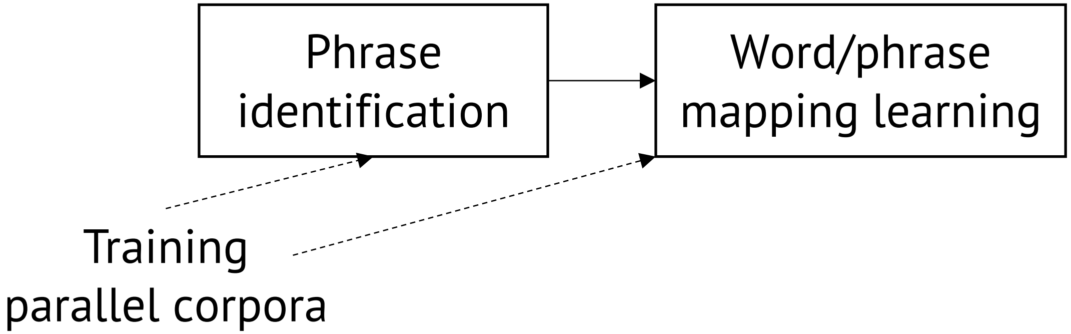
- Over-translation: unnecessary repeats of words/phrases in the translation

English (original)	Chinese (translated)
U have to admit that something *can never be changed*, live with it or break with it!	你必须承认，有些东西是永远无法改变的，无法改变的， 无法改变的，无法改变的!

Example of over-translation

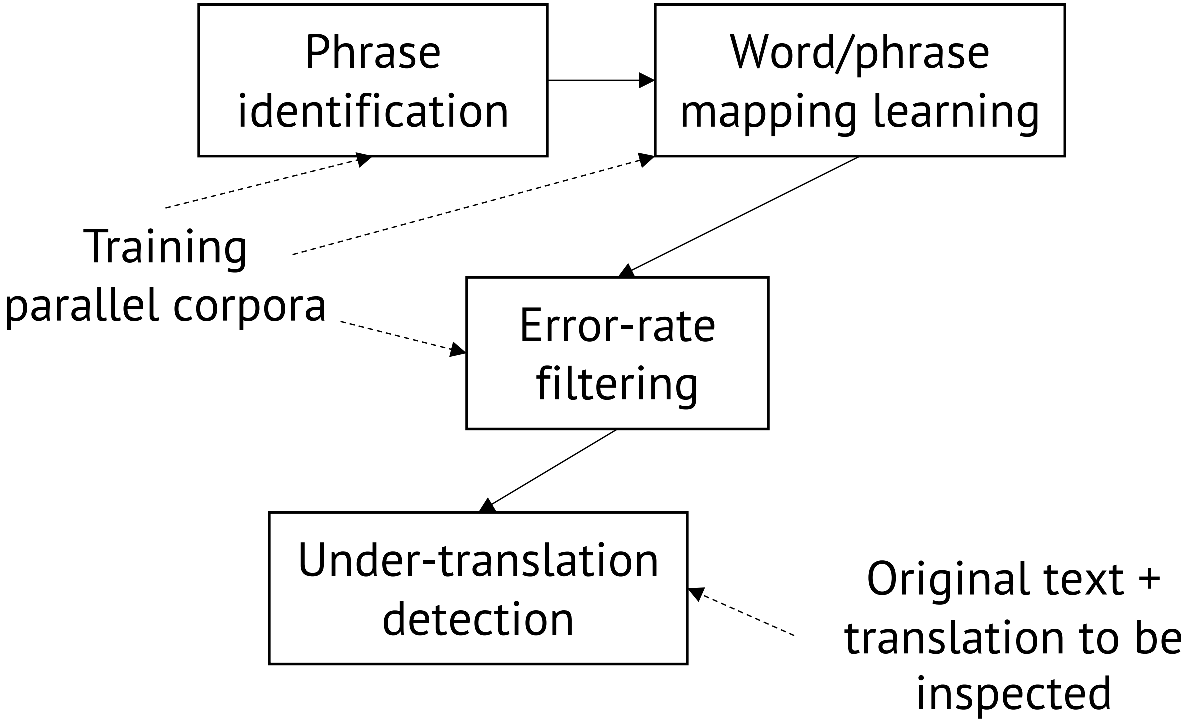
Overview of Violation Detection

- First step: build mappings between bilingual words/phrases using training parallel corpora



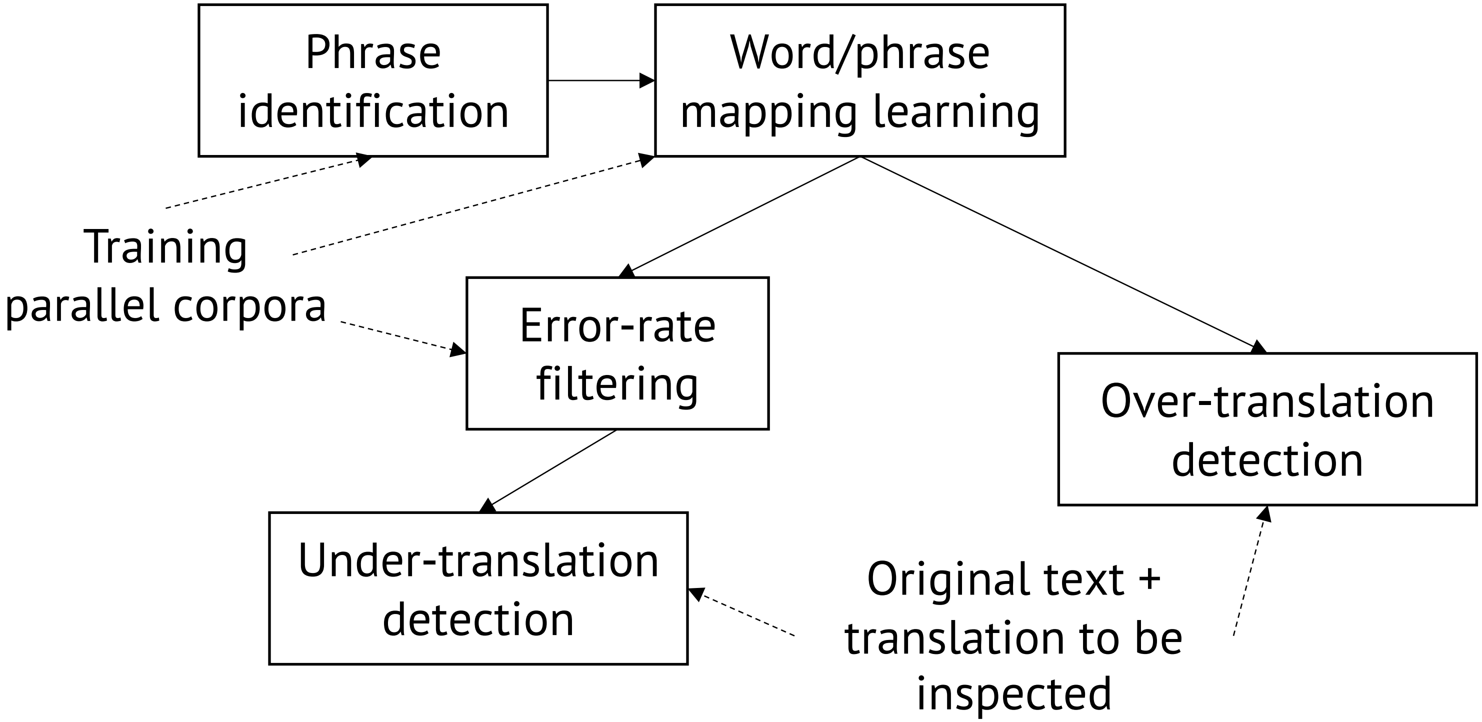
Overview of Violation Detection

- Under-translation detection: check the existence of word/phrase translations w.r.t. mappings
  - Need to consider implicit translations


Phrase identification
Word/phrase mapping learning
Training parallel corpora
Error-rate filtering
Under-translation detection
Original text + translation to be inspected

Overview of Violation Detection

- Over-translation detection: compare the occurrences of words/phrases in the original text and translation

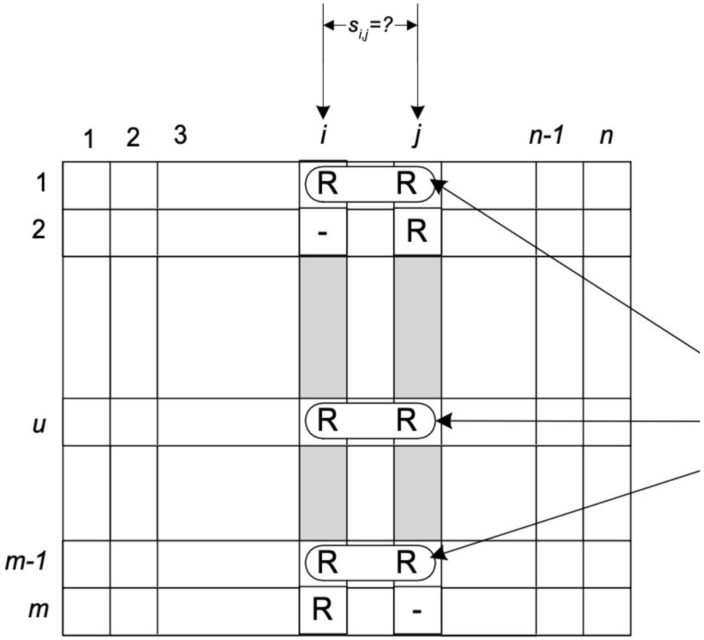

Phrase identification
Word/phrase mapping learning
Training parallel corpora
Error-rate filtering
Over-translation detection
Under-translation detection
Original text + translation to be inspected

Bilingual Mapping Building: Phrase Identification

- Necessary because phrases can convey different meanings from just their comprising words
- Intuitive way: consider all frequently-occurring continuous word sequences with length <= *k*
    - $w_1w_2w_3w_4w_5$… -> <w_1,w_2,w_3> <w_2,w_3,w_4> <$w_3w_4w_5$>…
    - But phrases can have variations
- Our approach: consider frequently-occurring word pairs that are <= *k* away from each other
    - $w_1w_2w_3w_4w_5$… -> <w_1,w_2> <w_1,w_3> <w_1w_4> <w_2,w_3> <w_2,w_4> <w_2,w_5>… (*k* = 3)
    - For both efficiency and robustness

Bilingual Mapping Building: Mapping Learning

- Item-based Collaborative Filtering
  - User rating matrix -> item recommendations
  - Similar items should have similar rating distributions



- 1,2,..,n represent items
- 1,2,…,m represent users

Credit: Sarwar, Badrul Munir, George Karypis, Joseph A. Konstan, and John Riedl. "Item-based Collaborative Filtering Recommendation Algorithms." WWW 2001.

Bilingual Mapping Building: Mapping Learning

- Item-based Collaborative Filtering
  - Item -> each word/phrase in the source/destination languages
  - User -> each bilingual sentence pair
  - Rating -> whether the word/phrase appears in the sentence pair (of the corresponding language)
  - Similarity -> Cosine similarity of rating vectors

$$M_{k,w} = \begin{cases} 1 & \text{if } w \text{ appears in } P_s^k \text{ or } P_d^k \\ 0 & \text{otherwise} \end{cases}$$

$$R_{w_s,w_d} = \frac{\overrightarrow{M_{\cdot,w_s}} \cdot \overrightarrow{M_{\cdot,w_d}}}{||\overrightarrow{M_{\cdot,w_s}}||_2 \cdot ||\overrightarrow{M_{\cdot,w_d}}||_2} = \frac{\sum_k M_{k,w_s} M_{k,w_d}}{\sqrt{\sum_k M_{k,w_s}^2}\sqrt{\sum_k M_{k,w_d}^2}}$$

Under-translation Detection

- Check the existence of each word/phrase translation w.r.t. mappings

Chinese (original)	English (translated)	English (reference)
三姑给你的红包给你妈妈了	Third Aunt gave you a red envelope.	Third Aunt gave your red envelope to your *mother*.

Origin	# 1	# 2	# 3	# 4	# 5
妈妈	mother	mom	mum	mama	mommy

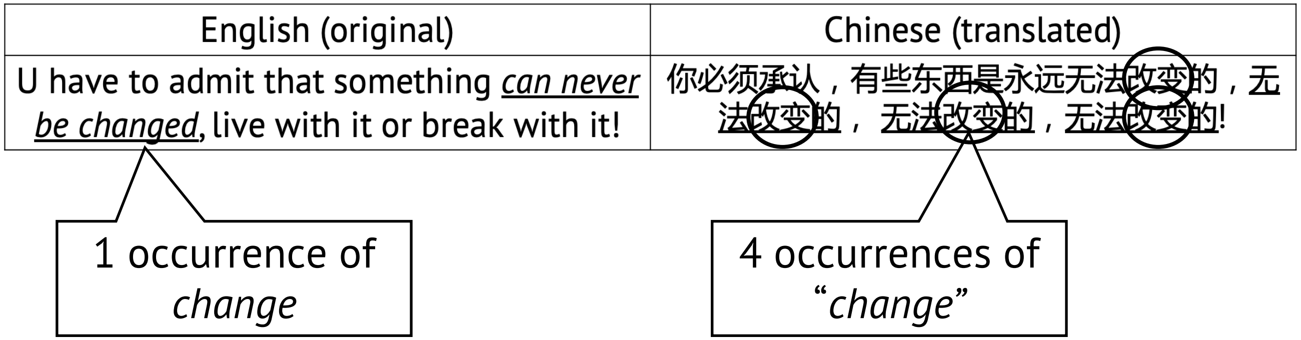
- Caveat: implicit translations
    - Some words/phrases might not need to appear in the translation text

Under-translation Detection: Handling Implicit Translations

- Error-rate filtering
  - A word/phrase causes too many translation failures -> Likely the word/phrase does not need to be explicitly translated
- $e_w = \#_w^{err}$ / $\#_w$ for each word/phrase w
  - Calculated on the training corpora
- A pre-defined threshold from experiments
  - $e_w < 0.2$ in our case

Over-translation Detection

- Find duplicate words/phrases in the translation
  - Not sufficient evidence of over-translation
- Reverse-lookup duplicated words/phrases w.r.t. mappings
- Is # of corresponding words/phrases < duplicated translation occurrences?



English (original)	Chinese (translated)
U have to admit that something *can never be changed*, live with it or break with it!	你必须承认，有些东西是永远无法改变的，无法改变的， 无法改变的，无法改变的!

1 occurrence of *change*

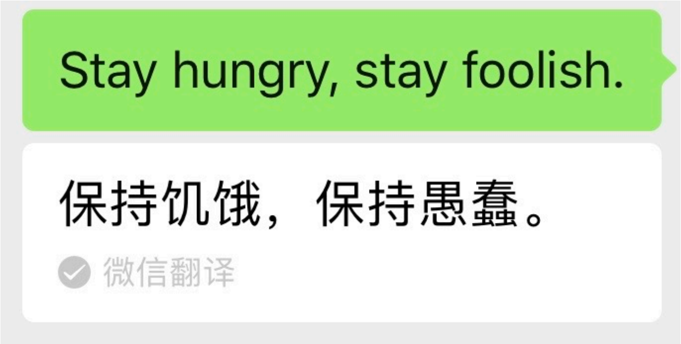
4 occurrences of “*change*”

Algorithm Effectiveness Evaluation

- 4 manually labeled datasets
  - Real-world translation tasks + corresponding translations with under-/over-translation
  - News and oral sentences between English and Chinese
- 2 alternative algorithms for comparison
  - Generic dictionary lookup
  - Word-alignment from SMT
- Highest F-measures in all tasks

Experience of Deployment

- Deployed on WeChat, a messenger app with over *one billion* monthly active users worldwide
  - Message translation function, powered by a proprietary NMT system
- Process about *12 million* translation tasks daily

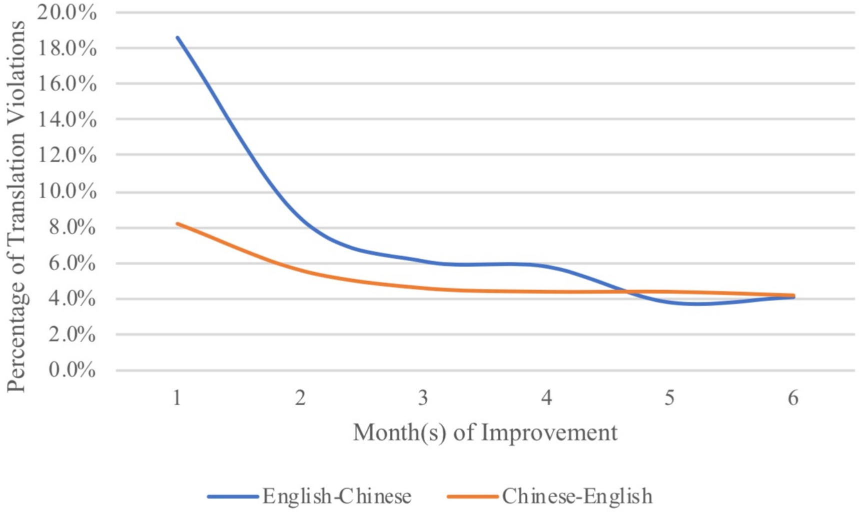


Stay hungry, stay foolish.

保持饥饿，保持愚蠢。

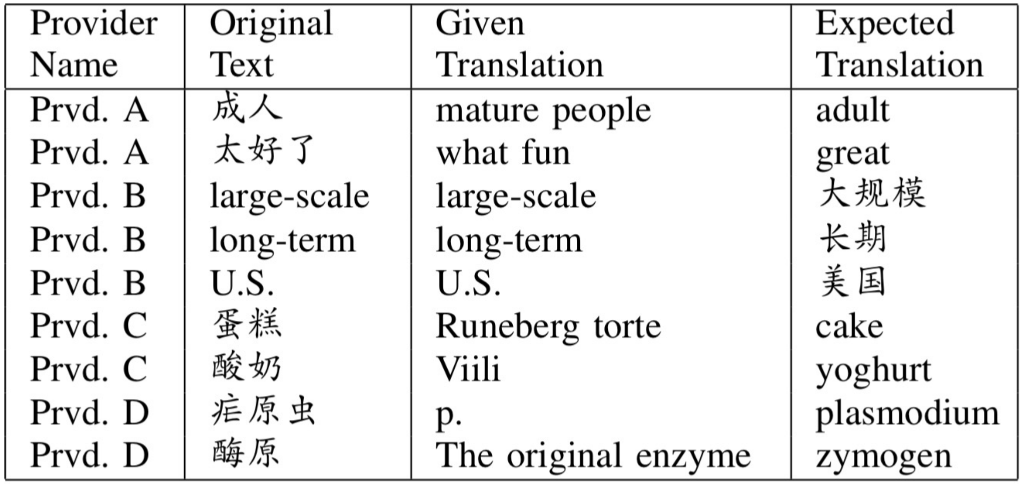
Experience of Deployment

- Fully rolled out in the production environment
  - Reveal issues undetected by in-house testing
  - Handle failures instantly through alternative models
  - Monitor the performance of newly-deployed models
- Lead to significant drop of two types of violations



Experience of Deployment

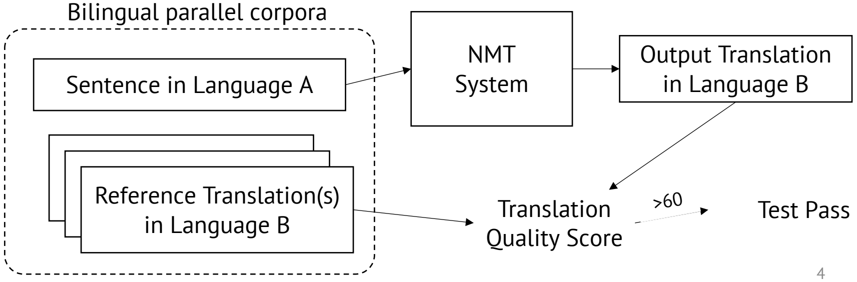
- Help build an in-house test set for regular development
  - 130,000 English and 180,000 Chinese words/phrases
  - Reveal design/implementation/training data defects in both ours and competing NMT systems



Provider Name	Original Text	Given Translation	Expected Translation
Prvd. A	成人	mature people	adult
Prvd. A	太好了	what fun	great
Prvd. B	large-scale	large-scale	大规模
Prvd. B	long-term	long-term	长期
Prvd. B	U.S.	U.S.	美国
Prvd. C	蛋糕	Runeberg torte	cake
Prvd. C	酸奶	Viili	yoghurt
Prvd. D	疟原虫	p.	plasmodium
Prvd. D	酶原	The original enzyme	zymogen

NMT Quality Assurance: Common Practice

- *Reference-based* black-box system testing
  - Performed during in-house development
  - Evaluate on human-made bilingual parallel corpora
  - Calculate and observe translation quality indicators (e.g., BLEU scores)



Bilingual parallel corpora

Sentence in Language A

Reference Translation(s) in Language B

NMT System

Output Translation in Language B

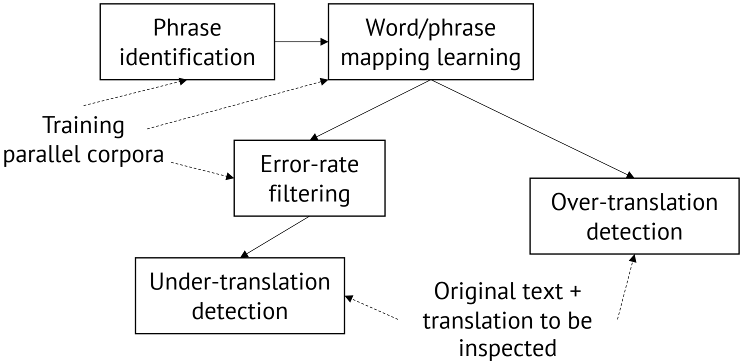
Translation Quality Score

>60

Test Pass

Overview of Violation Detection

- Over-translation detection: compare the occurrences of words/phrases in the original text and translation



Phrase identification

Word/phrase mapping learning

Training parallel corpora

Error-rate filtering

Over-translation detection

Under-translation detection

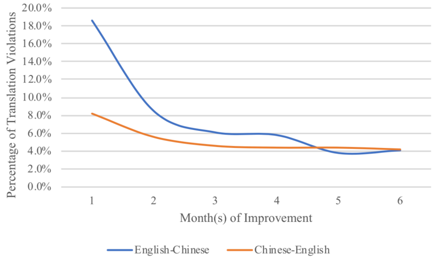
Original text + translation to be inspected

Algorithm Effectiveness Evaluation

- 4 manually labeled datasets
  - Real-world translation tasks + corresponding translations with under-/over-translation
  - News and oral sentences between English and Chinese
- 2 alternative algorithms for comparison
  - Generic dictionary lookup
  - Word-alignment from SMT
- Highest F-measures in all tasks

Experience of Deployment

- Fully rolled out in the production environment
  - Reveal issues undetected by in-house testing
  - Handle failures instantly through alternative models
  - Monitor the performance of newly-deployed models
- Lead to significant drop of two types of violations



Thanks!

The authors from Illinois CS were supported in part by NSF under grants no. CNS-1513939, CNS-1564274, CCF-1816615, and by a 3M Foundation Fellowship.