

A sunset over a beach with silhouettes of people in the foreground. The sun is a bright yellow circle in the center of the sky, which transitions from orange to a darker blue. The beach is dark, and the water reflects the sunset colors. Silhouettes of people are visible in the foreground, some standing and some sitting.

# 3-step Parallel Corpus Cleaning using Monolingual Crowd Workers

Toshiaki Nakazawa, Sadao Kurohashi  
(Kyoto University)

Hayato Kobayashi, Hiroki Ishikawa  
Manabu Sassano

(Yahoo Japan Corporation)

20/05/2015@PACLING2015

# Parallel Corpora

- Essential resources for almost all MT systems
- The **quality and quantity** greatly affect the translation quality
- Can be automatically constructed from existing resources
  - Europarl, patent families, Wikipedia...
- Need to **manually construct** it for domains which do not have enough existing resources

# Quality of Parallel Corpus

- Translation flaws are inevitable even though the professionals translate
  - *Homer nods* (弘法にも筆の誤り in Japanese)
- The number of flaws might be reduced by reviewing the whole corpus, but impossible
  - The size of the parallel corpus is usually very big
  - Very costly if we ask professionals to modify

# This Work

- Detect and edit the translation flaws in the existing manually-translated parallel corpus in effective and cheap way
- Use crowdsourcing in 3-steps
  1. Fluency Judgement
  2. Edit of Unnatural Sentences
  3. Verification of Edits
- The workers can be monolingual



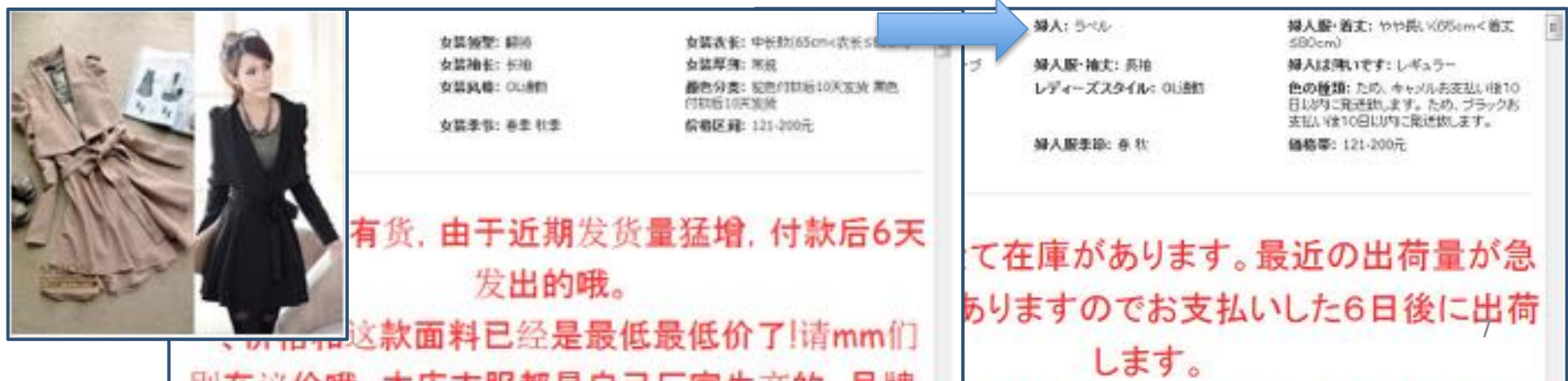
# Outline

- Motivation
- Brief introduction of collaborative research between Yahoo Japan and Kyoto University
- 3-step parallel corpus cleaning
- Experiments
  - Parallel corpus cleaning
  - Translation
- Conclusion and Future Work

# **BRIEF INTRODUCTION OF COLLABORATIVE RESEARCH**

# Collaborative Research Between Yahoo Japan and Kyoto University

- Goal: Improve the Chinese-to-Japanese translation for E-commerce site
- Task:
  - Develop a corpus-based MT system
  - Construct a parallel corpus of EC-site, especially fashion domain



The image shows a side-by-side comparison of a product page in Chinese and Japanese. On the left is the original Chinese page, and on the right is the translated Japanese page. A blue arrow points from the Chinese text to the Japanese text. The Chinese page features a model wearing a black dress, a list of specifications, and promotional text. The Japanese page shows the translated version of the same content.

Chinese Text	Japanese Text
女装类型: 裙装	婦人服: ワンピース
女装袖长: 长袖	婦人服・袖丈: 長袖
女装风格: OL通勤	レディーススタイル: OL通勤
女装季节: 春季 秋季	婦人服季節: 春秋
女装衣长: 中长款(65cm<衣长≤85cm)	婦人服・着丈: やや長い(65cm<着丈≤80cm)
女装厚薄: 常规	婦人は薄いです: レギュラー
颜色分类: 棕色(付款后10天发货 黑色 付款后10天发货)	色の種類: ため、キャンセルは支払い後10日以内に対応致します。ため、ブラックお支払い後10日以内に対応致します。
价格区间: 121-200元	価格帯: 121-200円

有货, 由于近期发货量猛增, 付款后6天发出的哦。

在庫があります。最近の出荷量が急増しておりますのでお支払いした6日後に出荷致します。

# Fashion-domain EC-site Parallel Corpus

[FDEC Corpus]

- 1.2M sentences (Zh: 6.3M, Ja: 8.7M words)
- Manually translated from fashion item pages of Chinese EC-site (taobao) into Japanese
- Most of the sentences were translated by **Chinese native speakers** (through the translation company)
  - Found many translation flaws in the Japanese translations



# Mother-tongue Principle

*“A translator should, as far as possible, translate into his own mother tongue or into a language of which he or she has a mastery equal to that of his or her mother tongue.”*

Recommendation on the Legal Protection of Translators and Translations and the Practical Means to improve the Status of Translators, UNESCO, 22 Nov. 1976

[http://portal.unesco.org/en/ev.php-URL\\_ID=13089&URL\\_DO=DO\\_TOPIC&URL\\_SECTION=201.html](http://portal.unesco.org/en/ev.php-URL_ID=13089&URL_DO=DO_TOPIC&URL_SECTION=201.html)

# Source Natives vs. Target Natives

	Source Natives	Target Natives
Background knowledge about the input sentence	High	Medium/Low
Fluency and grammatical correctness of the output sentence	Medium/Low	High

- Pros and cons for source and target native speakers
- Target natives for translation modification [Albrecht+, 2009]

# Examples of Translation Flaws

- Insertion

Ja: 随意にに1種類だけ注文

*(order one type at at your own will)*

- Remaining Chinese character

Ja: 元気あふれるという効果があります

Ref: 元気あふれるという効果があります

气 氣

Hanzi Kanji

- Unnatural

Ja: お手入れの時、電源を切れ、プラグを抜いてください。

*(when cleaning, turn the power off, please pull out the plug)*

# Other Translation Flaws

- Omission (not translated)

Zh: 看看有没有其他合适的商品

Ja: 看看有没有その他合適的商品

- Mistranslation

Zh: 加湿器功能 (*functions of humidifier*)

Ja: 除湿器の機能 (*functions of dehumidifier*)

Zh: 木耳



wood skirt with wood ear mushroom?

our framework **cannot** fix these kinds of flaws

**3-STEP**

**PARALLEL CORPUS CLEANING**

# 3-steps of Cleaning

## 1. Fluency Judgement

- detects the translation flaws

## 2. Edit of Unnatural Sentences

- edit the translated sentences

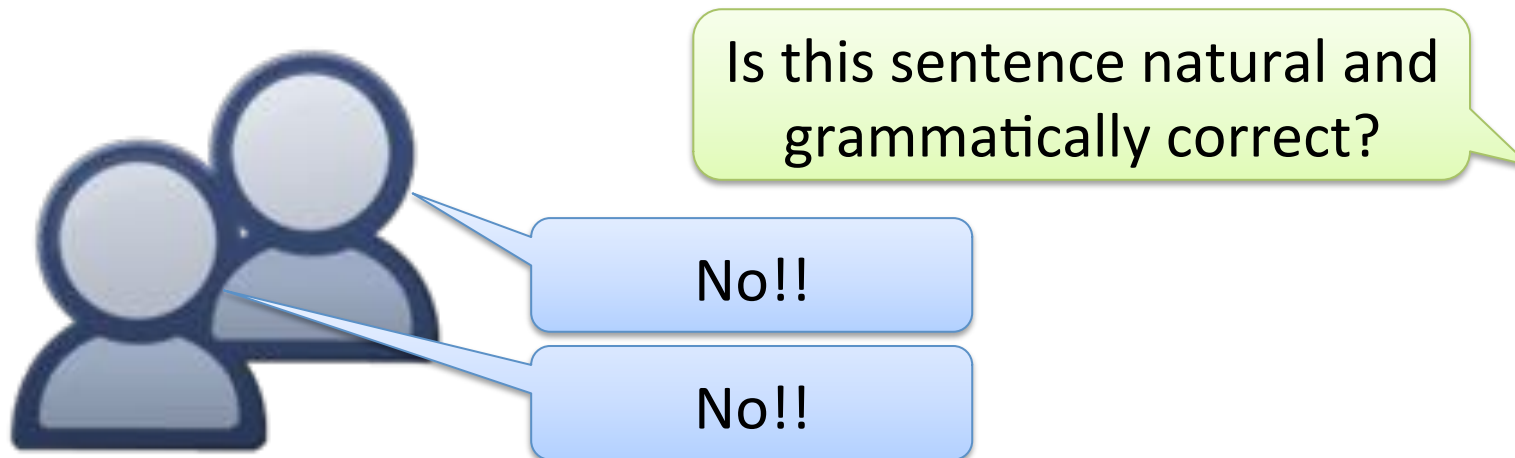
## 3. Verification of Edits

- check if the edited translation is better than the original one

# Step 1: Fluency Judgement

- Task: judge if the sentences are natural and grammatically correct
- Only showing the translated (target, Japanese) sentences

e.g. 随意にに1種類だけ注文



# Step 2: Edit of Unnatural Sentences

- Task: edit the unnatural translated sentences
- only showing the translated sentences, or show the source sentence as well for the reference

e.g. 随意にに1種類だけ注文（随便拍下一种）



Please modify this sentence to be natural and grammatically correct

随意に1種類だけ注文

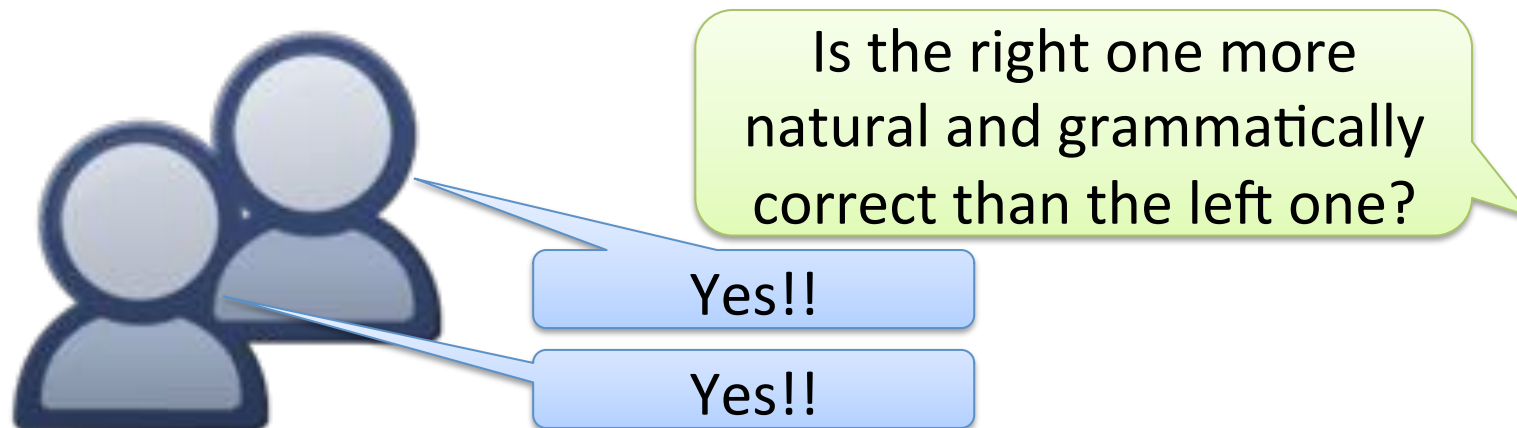
Nothing to modify!



## Step 3: Verification of Edits

- Task: judge if the edited translation is better than the original one
- This step is important to further improve the quality of the outcome because the edits are not necessarily correct

e.g. 随意にに1種類だけ注文 vs. 随意に1種類だけ注文



# **CORPUS CLEANING EXPERIMENTS**

# Crowdsourcing Service in the World





- Several styles of crowdsourcing tasks such as Yes/No questions and free writings
- The service is run in Japan; therefore most of the workers are Japanese
- Not able to select the workers by their abilities
- The workers in our experiments do not necessarily understand Chinese
  - perhaps almost all of them does not

# Step 1: Fluency Judgement

- 358,085 sentences from the FDEC corpus with length between 10 and 130 characters
- Only Japanese sentences are shown
- Asked 5 different workers for each question

# unnatural	5	4	3	2	1	0
# sents. ratio	13,056 (3.6%)	35,048 (9.8%)	60,200 (16.8%)	83,150 (23.2%)	93,187 (26.0%)	73,444 (20.5%)

**30%!**

# Step 2: Edit of Unnatural Sentences

- 47,420 sentences which were judged as unnatural by 4 or more workers in Step 1
- Original Chinese sentence is also shown
- Asked 3 different workers for each question

# edits	3	2	1	0
# sents.	3,755	12,498	18,289	12,878
ratio	(7.9%)	(26.4%)	(38.6%)	(27.2%)

## Step 3: Verification of Edits

- 54,550 edits which were generated in Step 2
- Original Chinese sentence is also shown
- Asked 5 different workers for each question

# better	5	4	3	2	1	0
# sents. ratio	25,053 (45.9%)	16,478 (30.2%)	7,706 (14.1%)	3,338 (6.1%)	1,462 (2.7%)	513 (0.9%)

# Translation Experiment

- Dataset: whole FDEC Corpus

# sentences	Original	Cleaned
Train	1,220,597	1,256,908
Dev	11,186	11,489
Test	11,200	11,495

- Cleaned: verified to be better by the majority
- Decoder: KyotoEBMT [Richardson+, 2014]
- Evaluation: BLEU



# Experimental Results

Train	Original	Cleaned	Cleaned	Cleaned
Dev	Original	Original	Cleaned	Cleaned
Test	Original	Original	Original	Cleaned
BLEU	21.39	<b>21.69</b>	21.34	21.12

- Corpus cleaning contributes to improve the translation quality!
- Cleaning the Dev and Test sets has bad effect on translation quality...

# Natural, but Incorrect/Unequal

- Reviewed 100 edits which are judged to be more natural than the original sentence by 5 workers
- Found 3 types of inequalities
  1. deletion of symbols (8 cases)
  2. omission (13 cases)
  3. mistranslation (5 cases)

see the proceedings for detailed examples

# Experimental Results

Train	Original	Cleaned	Cleaned	Cleaned
Dev	Original	Original	Cleaned	Cleaned
Test	Original	Original	Original	Cleaned
BLEU	21.39	<b>21.69</b>	21.34	21.12

- The inequalities have bad effect on the automatic evaluation scores because they suppose the content of the input and output are strictly equal

# Crowdsourcing Cost

- Cost for cleaning 6.8M words used in the experiments

	Professional*	Our Work
Fee	40 million JPY	2.6 million JPY
Time	1700 days	186 hours

\* These values are estimated from  
<http://www.editage.com>

# Conclusion and Future Work

- Proposed a framework of cleaning existing parallel corpora efficiently and cheaply
  - 3-step monolingual crowdsourcing
  - Improved the fluency of the sentences
- Future work
  - How to reduce the inequalities of the edits?
  - How to improve the correctness of the translation by monolingual workers?



Thank you!  
Terima kasih!