# Fluency-guided Cross-lingual Image Captioning

**Weiyu Lan**[1], Xirong Li[1], Jianfeng Dong[2]

[1] Renmin University of China

[2] Zhejiang University


# Image Captioning

A person holding a book with a bird sitting on the book.

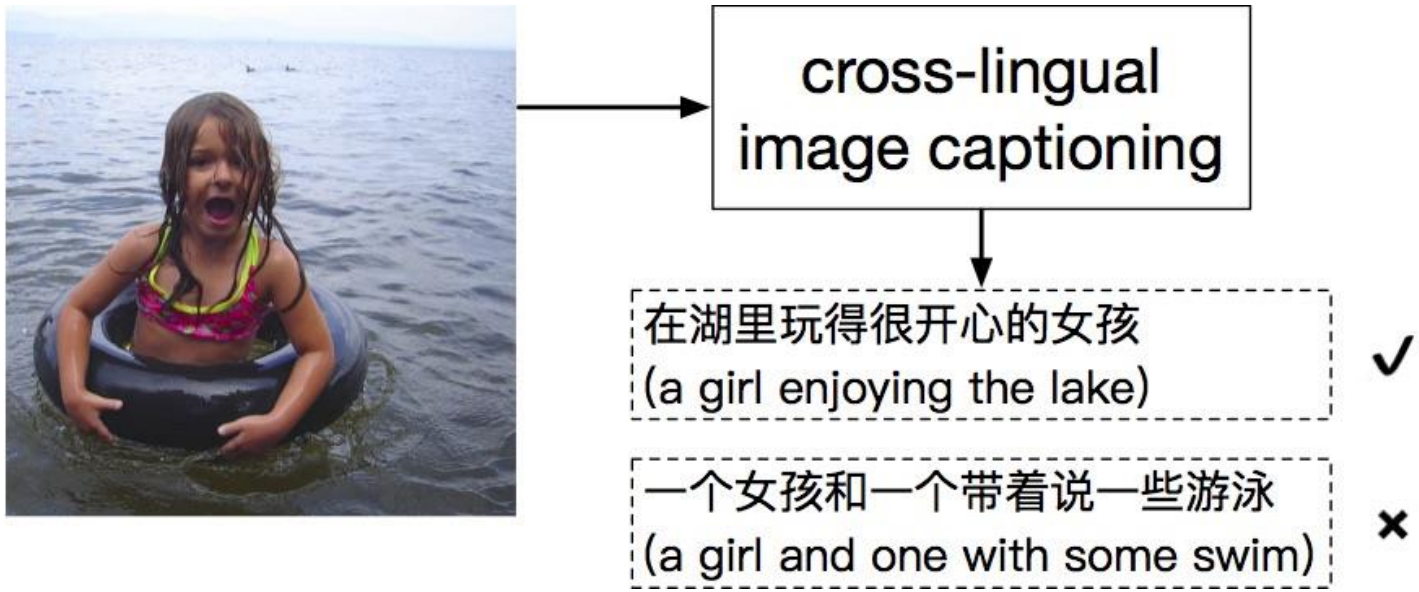
一个人拿着一本书，有一只小鸟站在上面

Una persona sostiene un libro con un pájaro que se sienta sobre el libro

책에 앉아 있는 새와 함께 책을 들고 있는 사람

# Cross lingual image captioning

- Goal: To generate **relevant** and **fluent** captions in a target language with minimal human effort



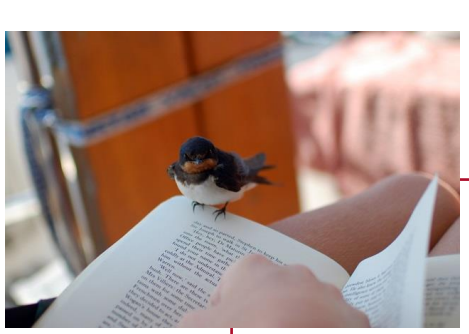
# Related Work

- Monolingual image captioning
  - Deep learning: Encoder + Decoder (CNN+RNN)

- Cross-lingual image captioning
  - Generate captions base on both image and captions in source language (Elliott et al., 2015)
  - Crowd sourcing to collect Japanese descriptions of the MSCOCO (Miyazaki and Shimizu, 2016)
  - **Machine Translation (Li, ICMR2016)**

# Related Work

- Cross-lingual image captioning
  - Machine Translation (Li, ICMR2016)



translated Chinese sentences

Chinese Captioning Model

一个人拿着一本书，
有一只小鸟站在上面 √

Captioning Model

A person holding a book with a bird sitting on the book

Machine translation

×

# Machine-translated sentences are not fluent

A person holding a book with a bird sitting on the book.

拿着一本书和一只鸟坐在书上的人

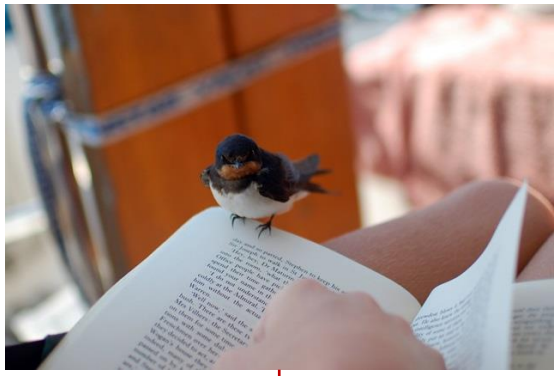
Una persona que sostiene un libro con un pájaro sentado en el libro.

책 한 권을 쥐고 한 사람 한 마리가 책 위에 앉아 있다.

?

# Our Approach

- **Fluency-guided** cross-lingual image captioning



(A person holding a book with a bird sitting on the book)
拿着一本书和一只鸟坐在书上的人 → Not Fluent

(A small bird sitting on top of an open book)
一只小鸟坐在一本打开的书上 → Fluent

Image

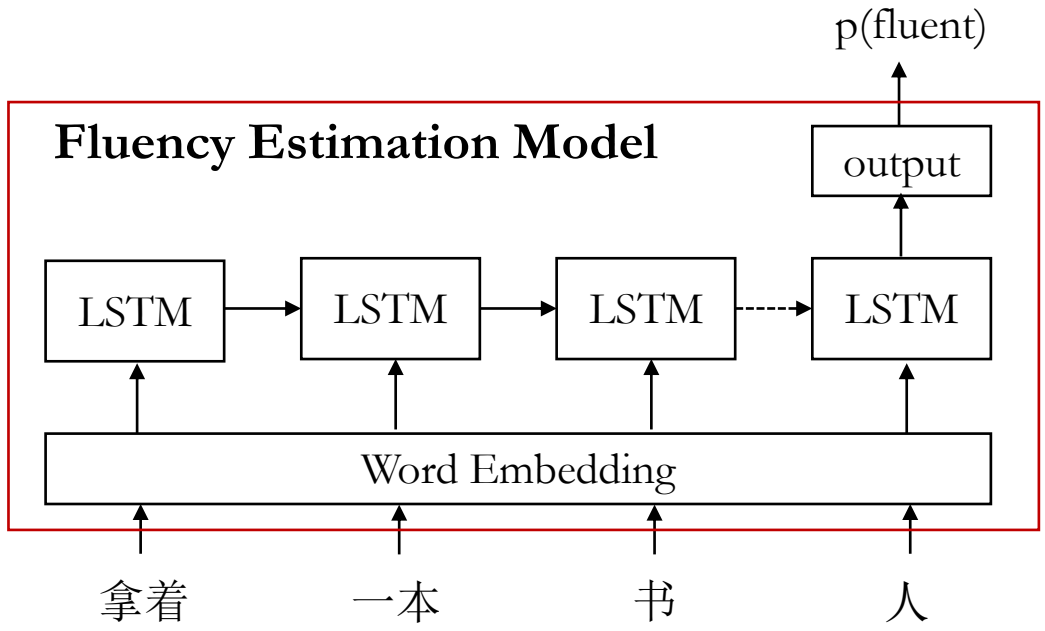
Sentence

Estimated Fluency

Captioning Model

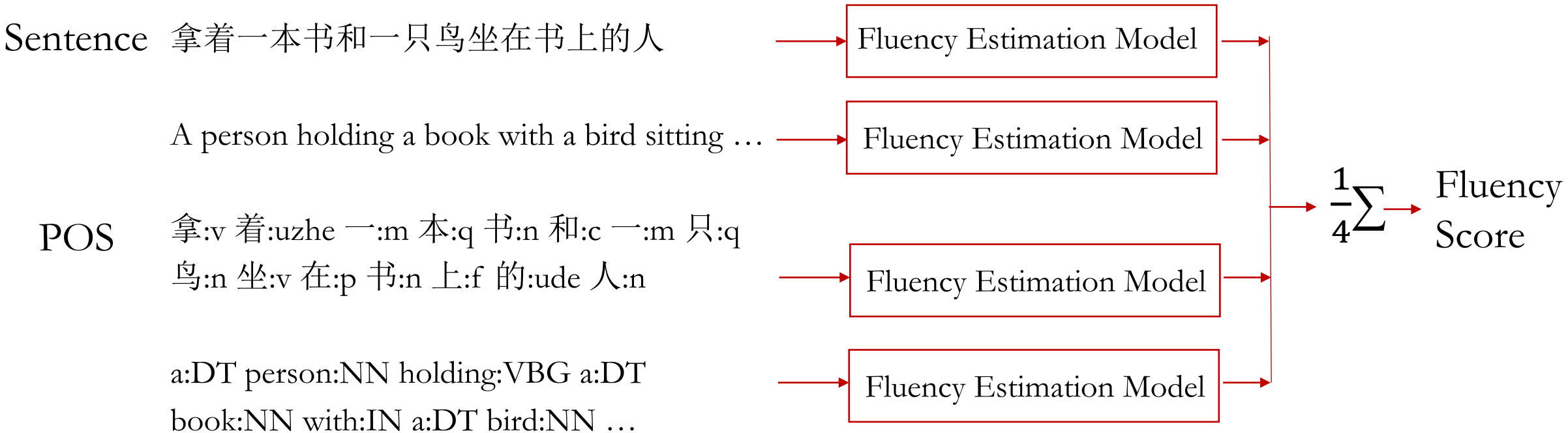
# Sentence Fluency Estimation

(A person holding a book
with a bird sitting on the book)
拿着一本书和一只鸟坐在
书上的人




- Binary classification
- Manual annotation
  - 8k sentences: fluent/not fluent
  - Less than 30% sentences are fluent
- LSTM based model

# A four-way LSTM based classifier

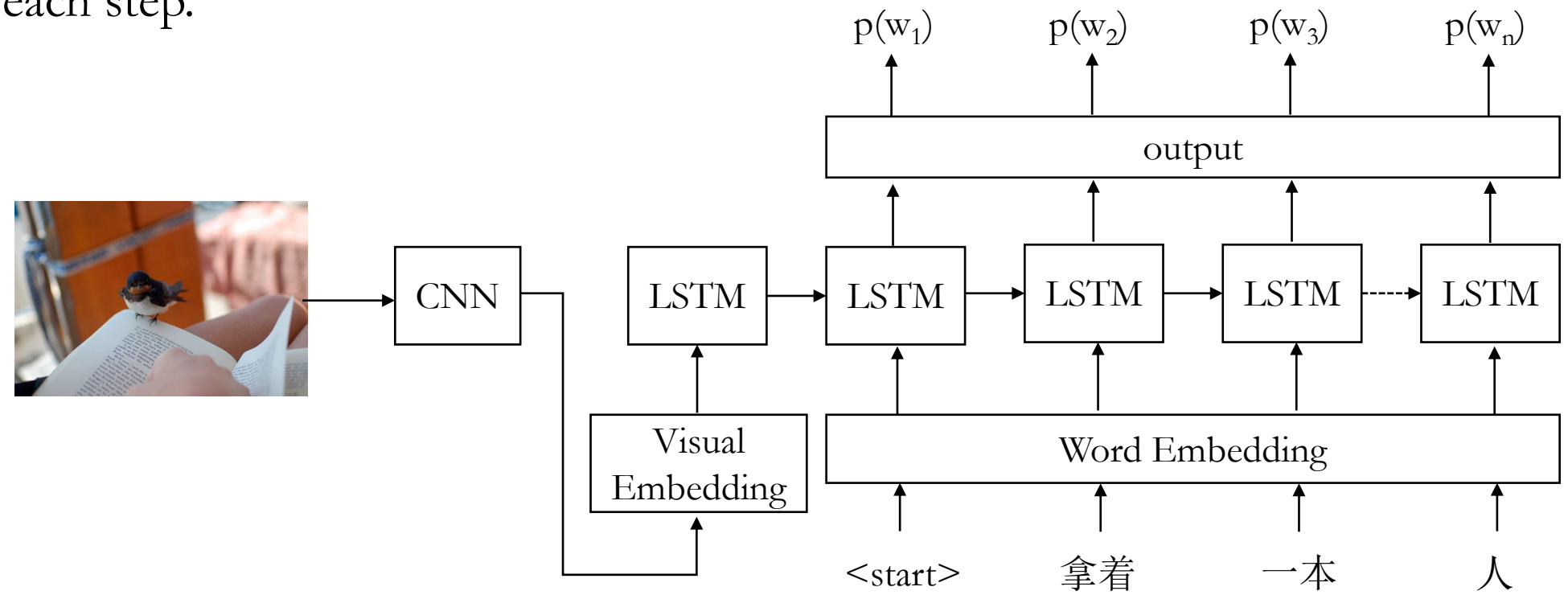


# Sentence Fluency Estimation Results

| English sentences | Chinese sentences | Estimated fluency scores |
|---|---|---|
| The two large elephants are standing in the grass | 两只大象正站在草地上 | 0.803 |
| The young man in the blue shirt is playing tennis | 穿蓝色衬衫的年轻人正在打网球 | 0.624 |
| A group of people riding skis in their bathing suits | 一群人在他们的沐浴骑滑雪服 | 0.117 |
| A sports arena under a dome with snow on it | 一个体育馆下一个圆顶下的雪在它 | 0.060 |

# Image Captioning Model

- CNN + RNN framework [Vinyals, CVPR2015]
- Training loss is the sum of the negative log likelihoods of the next correct word at each step.



# Fluency-Guided Training

# Fluency-Guided Training

## Strategy I: Fluency only

(A small bird sitting on top of an open book)
一只小鸟坐在一本打开的书上

0.9
Fluent

(A person holding a book with a bird sitting on the book)
拿着一本书和一只鸟坐在书上的人

0.2
Not Fluent

Captioning Model

# Fluency-Guided Training

## Strategy II: Rejection sampling

- Allow the sentences classified as not fluent to be used for training with a certain chance

(A small bird sitting on top of an open book)
一只小鸟坐在一本打开的书上

0.9
Fluent

(A person holding a book with a bird sitting on the book)
拿着一本书和一只鸟坐在书上的人

0.2
Not Fluent

Rejection Sampling

u~ U (0, 0.5)

Captioning Model

# Fluency-Guided Training

## Strategy III: Weighted loss

- Cost-sensitive learning

(A small bird sitting on top of an open book)

一只小鸟坐在一本打开的书上

(A person holding a book with a bird sitting on the book)

拿着一本书和一只鸟坐在书上的人

0.9
Fluent

0.2
Not Fluent

Captioning Model

$$Weighted\ batch\ loss = -\frac{1}{\mathrm{m}}(\log p(s_1) + 0.2 * \log p(s_2) + \cdots)$$

# Datasets and Experiments

# Developing test set

- Manually translating sentences in test set as ground truth
  - Providing both English sentence and corresponding image
  - To eliminate ambiguity and translate referring to the image

# Two Bilingual (English-Chinese) Datasets

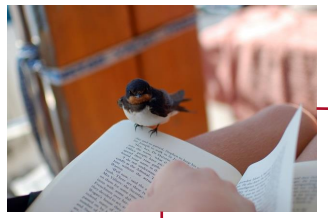
- Extending Flickr8k and Flickr30k to bilingual version (English + Chinese)

  Download: https://github.com/li-xirong/cross-lingual-cap

| | Flickr8k-cn | | | Flickr30k-cn | | |
|---|---|---|---|---|---|---|
| | Train | Validation | Test | Train | Validation | Test |
| Images | 6,000 | 1,000 | 1,000 | 29,783 | 1,000 | 1,000 |
| Machine-translated Chinese sentences | 30,000 | 5,000 | - | 148,915 | 5,000 | - |
| Human-translated Chinese sentences | - | - | 5,000 | - | - | 5,000 |
| Human-annotated Chinese sentences | 30,000 | 5,000 | 5,000 | - | - | - |

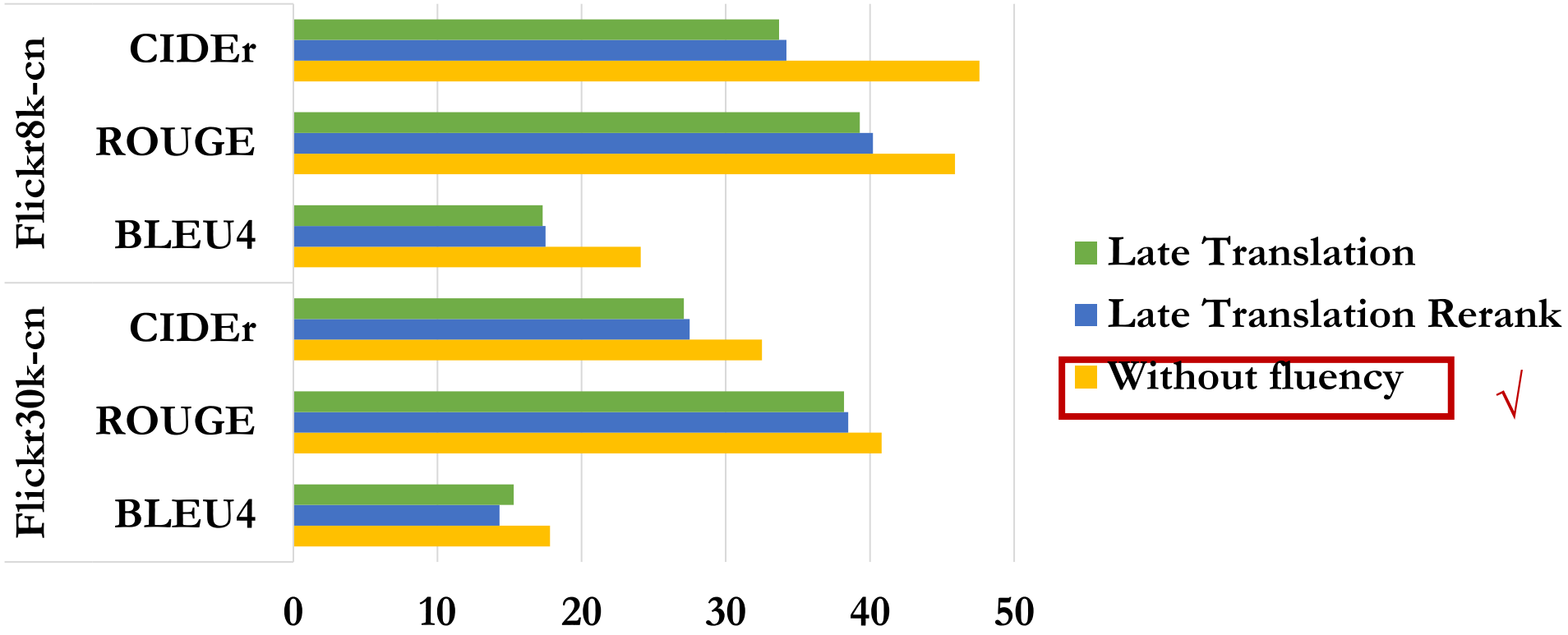
# Experiments

- Baselines:
  1. Late translation[Li, ICMR2016]
  2. Late translation rerank
  3. Without fluency
  4. Manual Flickr8k-cn
- Proposed approaches:
  1. Fluency-only
  2. Rejection sampling
  3. Weighted loss



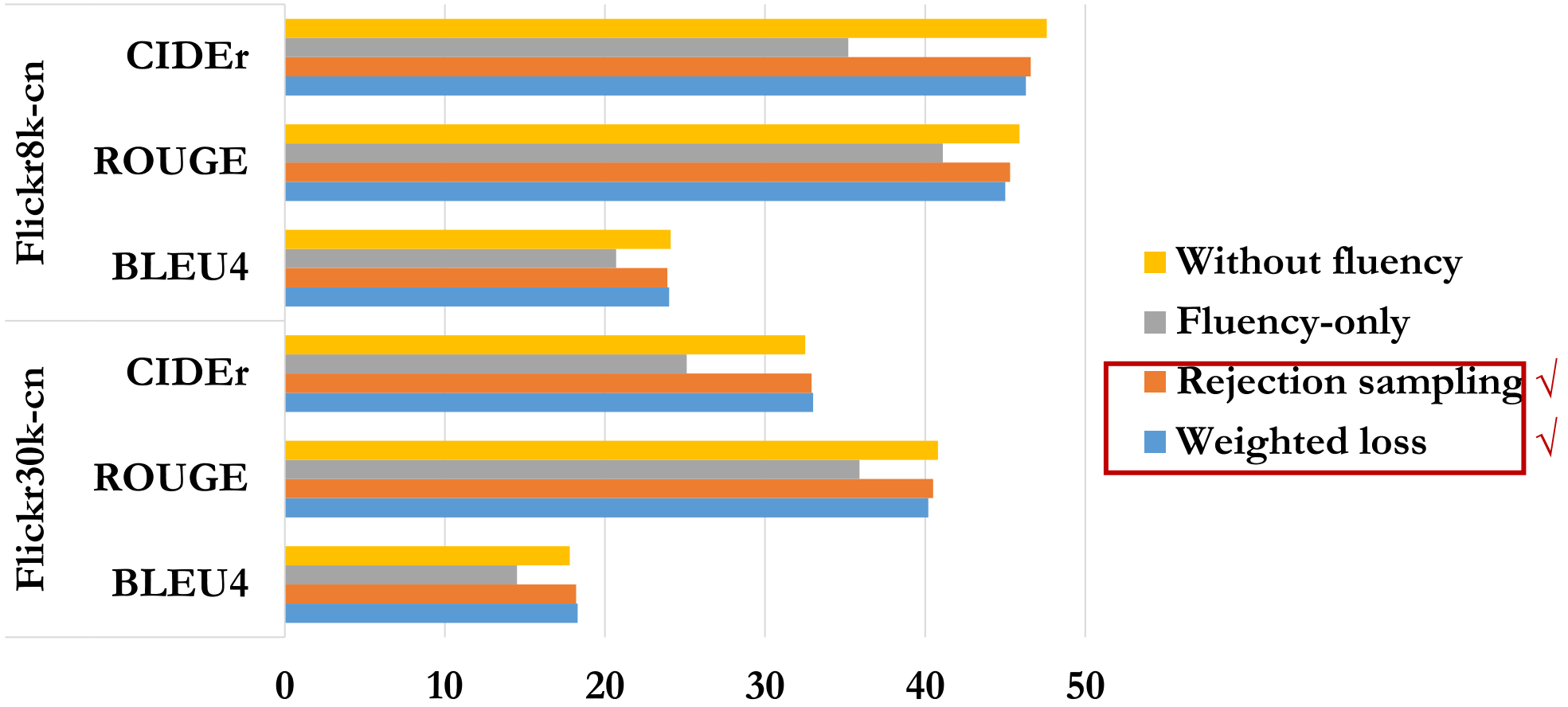
# Automatic Evaluation Results

- Late translation is not effective



# Automatic Evaluation Results

- **Rejection sampling** and **Weighted loss** are able to preserve relevant information



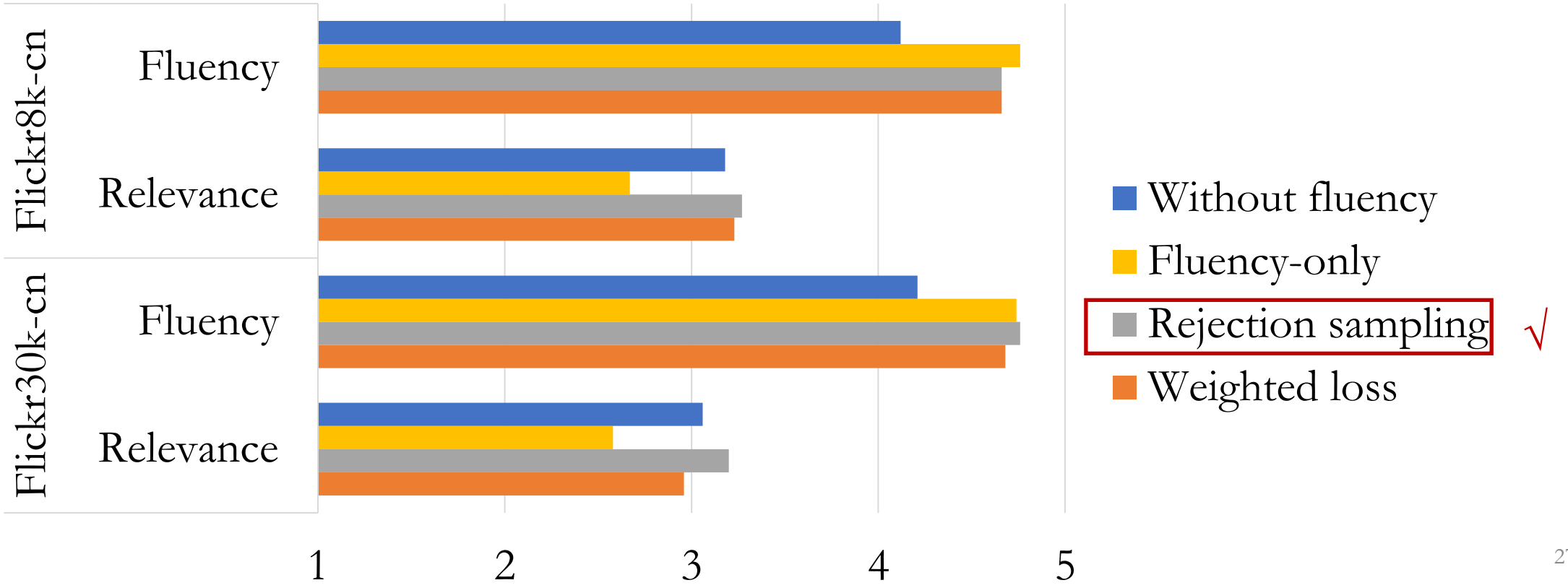
# Human Evaluation

- Automatic evaluation is insufficient to guarantee the overall fluency

- Annotators rate the sentences using a Likert scale of 1 to 5 (higher is better) in two aspects, namely **relevance** and **fluency**
  - Sentences generated by distinct approaches are shown together
  - Sentences randomly shuffled before presenting to the annotators

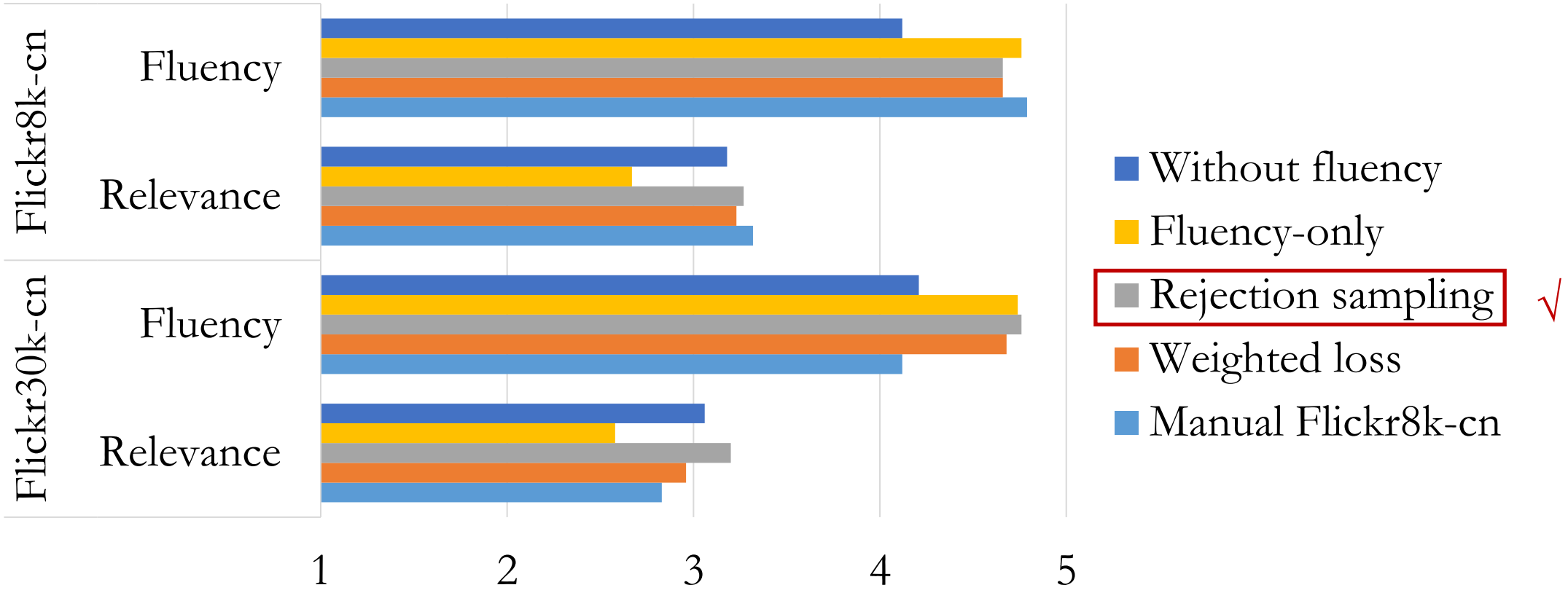
# Human Evaluation Results

- **Rejection sampling** achieves the best balance between relevance and fluency



# Human Evaluation Results

- **Rejection sampling** achieves the best balance between relevance and fluency, without the need of manual written Chinese captions.



# Conclusion

Fluency-guided framework

- Tackling cross-lingual image captioning with minimal manual annotation effort
- Capable of generating relevant and fluent captions in target language

https://github.com/li-xirong/cross-lingual-cap

bluey@ruc.edu.cn