Addressing the Under-translation Problem from the Entropy Perspective

Yang Zhao, Jiajun Zhang, Chengqing Zong, Zhongjun He, and Hua Wu

{yang.zhao, jjzhang, cqzong}@nlpr.ia.ac.cn, {hezhongjun,wu_hua}@baidu.com

Introduction

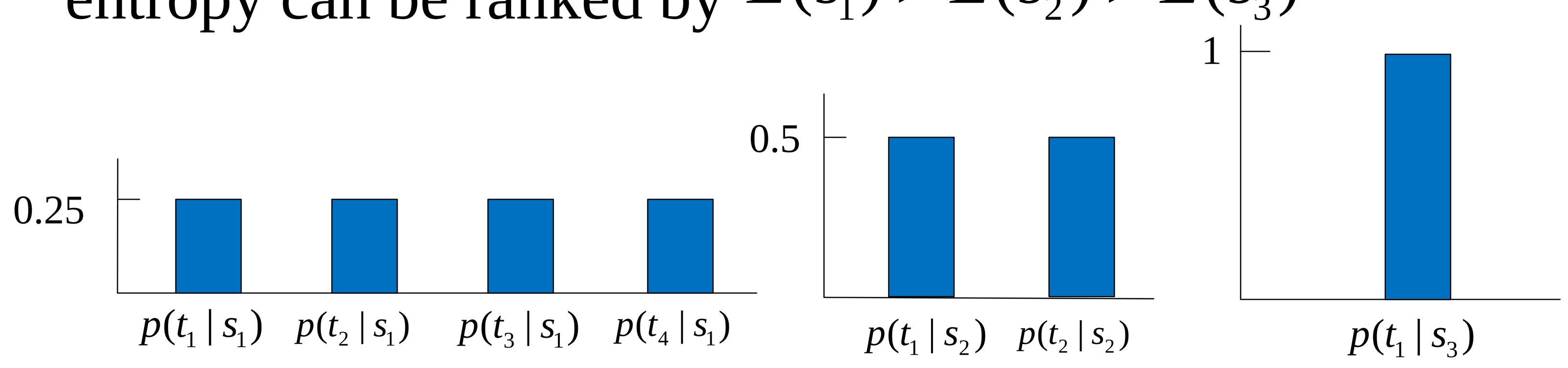
- **Under-translation:** some source words are mistakenly dropped by the neural machine translation (NMT) model.
- The current methods study the problem in the **model level**.
- **Our Contributions**
  - We find that source words with larger translation entropy are more likely to be dropped by the neural model.
  - We propose a coarse-to-fine framework to address the under-translation problem of high-entropy words.

Observation and Motivation

- **Definition (translation entropy):**

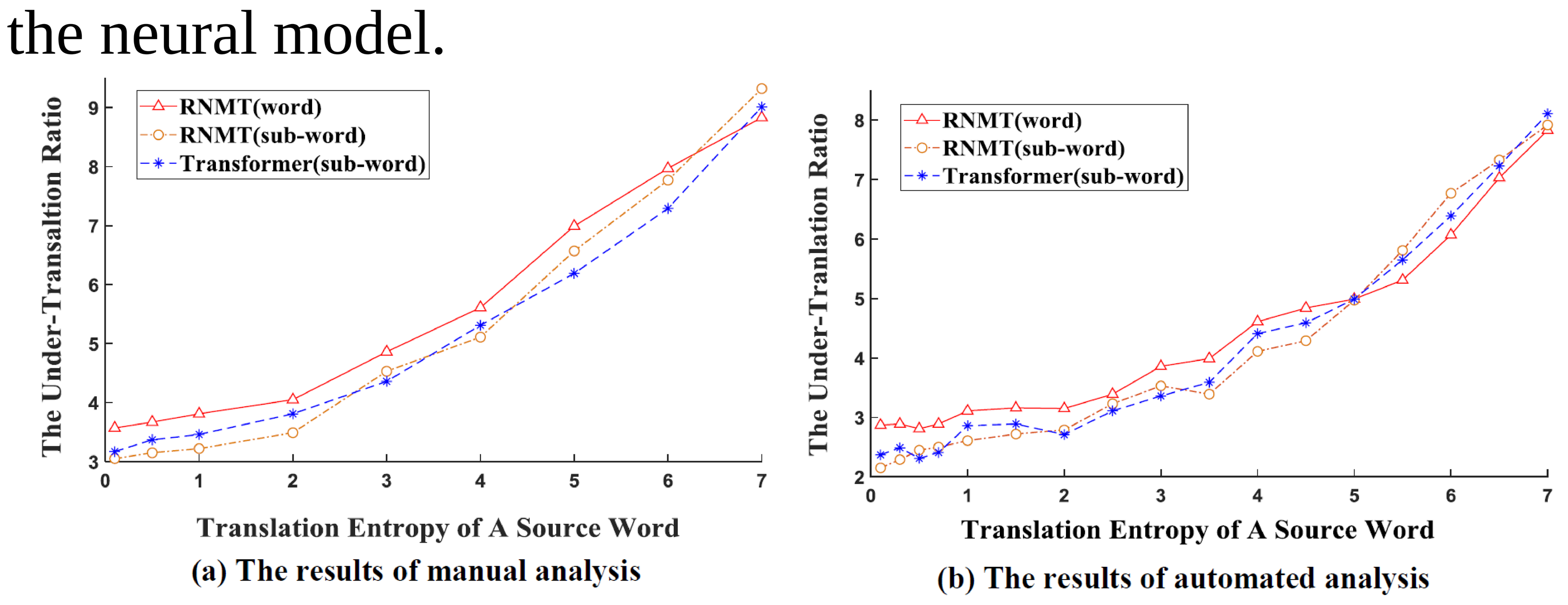
Assume a word S contains K candidate translations, each of which has a probability p_k, the translation entropy for this word can be calculated by $E(s)=-\sum_{k=1}^{K} p_k \log p_k$

- **Example**. For the following three words, the translation entropy can be ranked by $E(s_1)>E(s_2)>E(s_3)$



(a) word s_1 (b) word s_2 (c) word s_3

Observation: For a source word s, the larger its translation entropy is, the more likely this word is to be ignored by the neural model.



(a) The results of manual analysis (b) The results of automated analysis

High-entropy Words

- If the translation entropy exceeds the predefined threshold, we treat this word as a high-entropy word.
- Our goal is to reduce the under-translation cases of these high-entropy words.

Method Description

We propose a **coarse-to-fine framework** to address this problem

- **Coarse-grained Phase**
  - we construct the **pseudo target sentences** to reduce the entropy.
  - **construction method:**

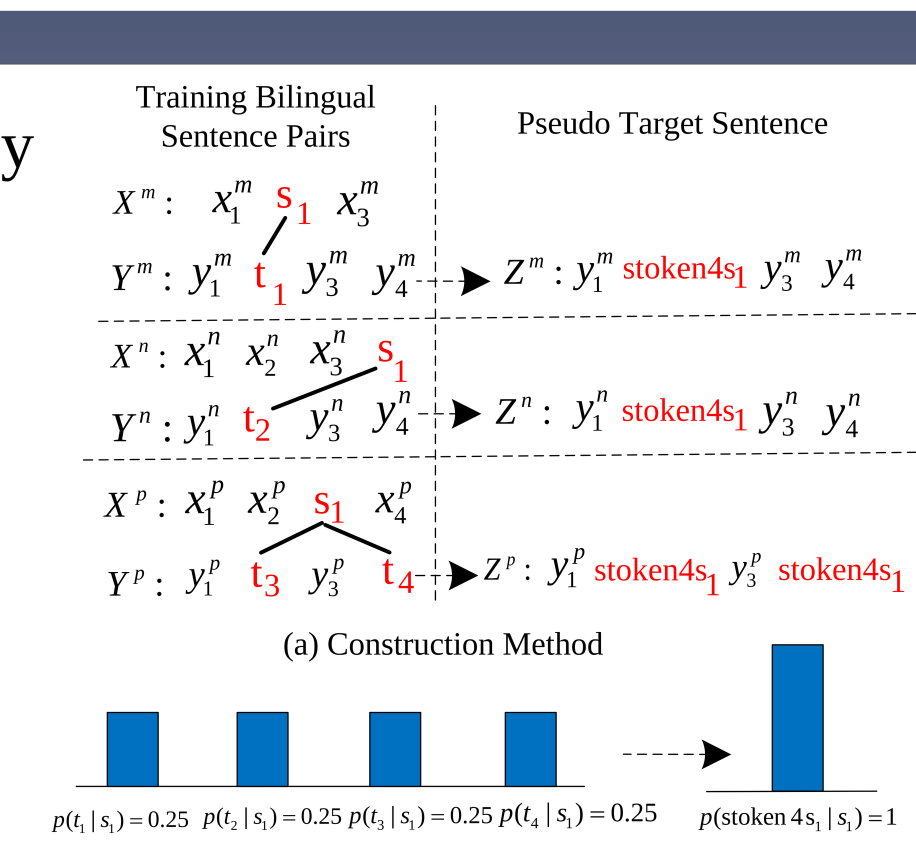
replacing the candidate translations of each high-entropy word with its respective special pseudo token.

Training data will change by

$$D_{xy}=\{X^{(n)},Y^{(n)}\}_{n=1}^{N}$$

$$D_{xyz}=\{X^{(n)},Y^{(n)},Z^{(n)}\}_{n=1}^{N}$$

- Z is the derived pseudo target sentence



- **Fine-grained Phase**

the derived pseudo sentences are utilized to improve the neural model.

- **Pre-training method**
- **Multitask method**

train neural model through two translation tasks:

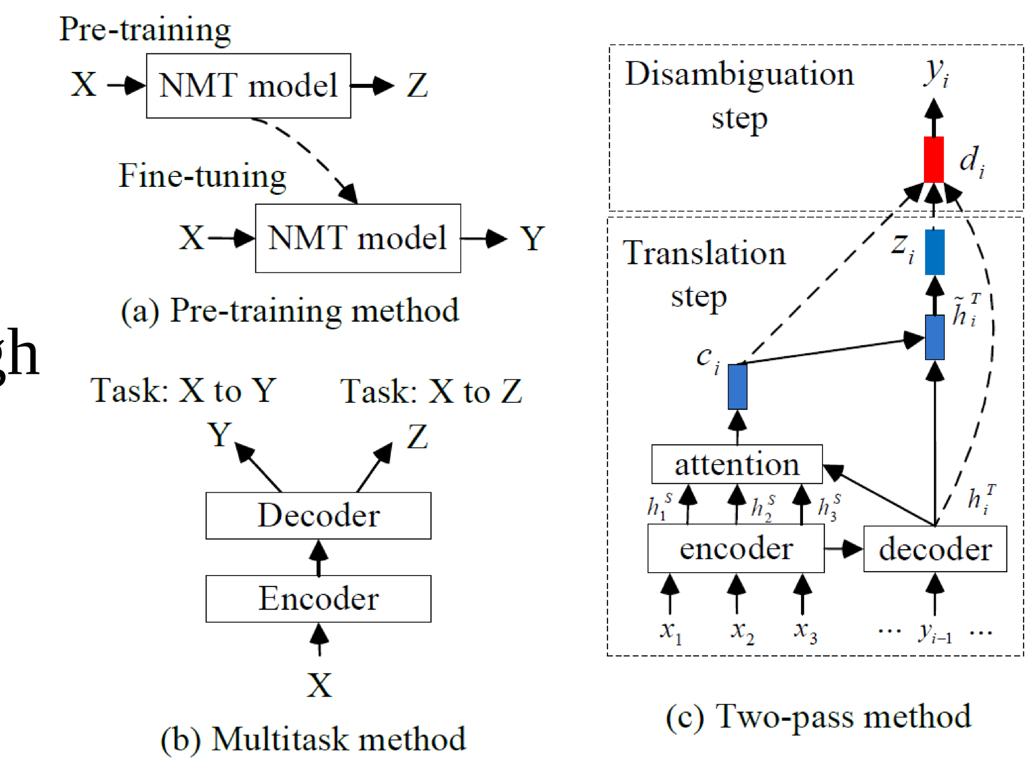
  - X to Y and X to Z

- **Two-pass Method**
- **Translation step:**

Translate X to Z with the standard NMT model

- **Disambiguation step:**

Transform the special token in Z to real target word in Y with a network



Experiments

#	Model	Units	03	04	05	06	08	Avg.
1	RNMT	word	41.01	42.94	40.31	40.57	30.96	39.16
2	RNMT+pre_train	word	41.53†	43.46*	40.41*	41.35†	31.17*	39.58
3	RNMT+multitask	word	41.99†	43.95†	40.84†	41.57†	31.42*	39.95
4	RNMT+two_pass	word	**42.37**†	**44.27**†	**41.58**†	**41.72**†	**31.91**†	**40.37**
5	RNMT	sub-word	43.96	44.74	42.46	43.01	32.53	41.34
6	RNMT+pre_train	sub-word	44.13	44.96*	42.61*	43.31*	32.77*	41.56
7	RNMT+multitask	sub-word	44.53†	45.17*	**43.39**†	43.85†	32.97*	41.98
8	RNMT+two_pass	sub-word	**44.86**†	**45.64**†	43.36†	**44.17**†	**33.44**†	**42.29**
9	RNMT+coverage	word	41.75	43.79	41.44	41.24	31.46	39.94
10	RNMT+coverage+multitask	word	42.66†	**44.54**†	42.07†	41.77†	32.03†	40.61
11	RNMT+coverage+two_pass	word	**42.94**†	44.52†	**42.53**†	**42.12**†	**32.23**†	**40.87**
12	Transformer	sub-word	45.80	47.77	46.90	46.90	34.61	44.40
13	Transformer+multitask	sub-word	**46.71**†	48.13*	47.41†	47.44†	34.98*	44.93
14	Transformer+two_pass	sub-word	46.64†	**48.29**†	**47.63**†	**47.51**†	**35.13**†	**45.04**

The BLEU points of CH-EN translation

Model	All	High-Entropy	Other
RNMT	5.02%(207)	8.05%(91)	3.88%(116)
+two_pass	4.10%(169)	4.86%(55)	3.81%(114)
Coverage	4.32%(178)	7.07%(80)	3.28%(98)
+two_pass	3.59%(148)	4.77%(54)	3.14%(94)
Transformer	4.63%(191)	7.60%(86)	3.51%(105)
+two_pass	3.78%(156)	4.69%(53)	3.44%(103)

The under-translation ratio (number) of different methods.

Model	All	High-Entropy
RNMT	45.54	51.34
RNMT+ multitask	43.39	47.18
RNMT+ two_pass	42.21	46.37
Transformer	44.31	49.47
Transformer +multitask	41.41	45.11
Transformer +two_pass	40.22	44.02

The alignment error rate of different methods.

Conclusion

1. We find that source words with larger translation entropy are more likely to be dropped.
2. We propose a coarse-to-fine framework to address this problem.
3. The experiments demonstrate that our method can sharply reduces the under-translation cases of these high-entropy words.

National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences

Baidu Inc.