陕西省复杂系统控制与智能信息处理重点实验室
Laboratory for Complex System Control and Intelligent Information Processing

Sentence-length Informed Method for Active Learning based Resource-poor Statistical Machine Translation

Speaker: Jinhua Du (杜金华)

Multi-language Information Interaction and Process Lab,
Xi'an University of Technology
西安理工大学
多语言信息交互与处理研究室

jhdu@xaut.edu.cn

西安理工大学 XI'AN UNIVERSITY OF TECHNOLOGY

Outline

- ✓Introduction
- ✓Sentence-length Informed Method
- ✓Experiments and Analysis
- ✓Conclusions and Future Work

jhdu@xaut.edu.cn

Introduction - Questions

- The large scale high quality parallel data is very crucial for good translation performance. However, it is not the case for many resource-poor language pairs.

- Many ways to alleviate this problem, such as collecting data from the Web, human translation, etc.
- The active learning (AL) framework also provides an effective way to facilitate the parallel data.

- The key issue in the active learning strategy is to choose rich-information sentences in order to maximize the value of human cost.

Introduction - Related work

In 2009, Haffari et al. firstly proposed a practical active learning framework for SMT where a number of high-quality parallel data are acquired from the large-scale monolingual data.

In 2010, Ambati et al. proposed an active crowd translation (ACT) paradigm where active learning and crowd-sourcing come together to enable automatic translation for low-resource language pairs.

In 2012, Bakhshaei and Khadivi applied a pool-based AL strategy to improve Farsi-English SMT system.

In 2013, we proposed a simple but effective sentence length control method to restrain the bias of the algorithm on short sentences.

Introduction – AL Framework for SMT

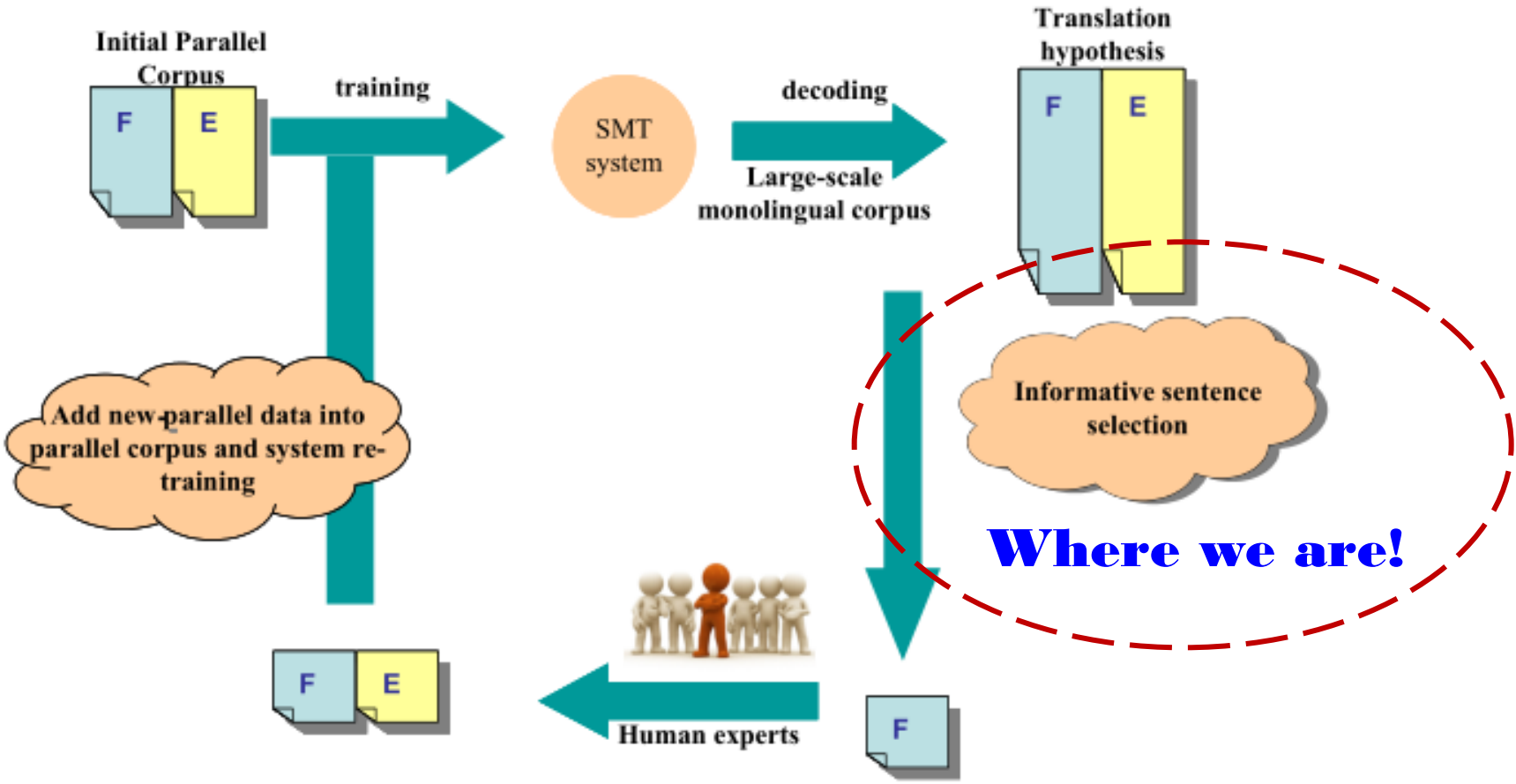


Fig.1. The workflow of active learning framework for SMT

jhdu@xaut.edu.cn

Introduction – General Algorithm Description

The active learning framework using translation units based algorithm in our work is as follows:

Algorithm 1 Modified AL-SMT

1: Given bilingual corpus L, and monolingual corpus U.
2: $M_{F \to E} = \textbf{train}(L)$
3: **for** $t = 1, 2, \ldots, N$ **do**
4: Generate "Phrase Set" and compute sentence scores
5: Select k sentences from U, and ask human experts for true translations.
6: Remove the k sentences from U, and add the k sentence pairs to L.
7: Update $M_{F \to E} = \textbf{train}(L)$
8: Evaluate the system performance on the test set.
9: **end for**

jhdu@xaut.edu.cn

Outline

- ✓Introduction
- ✓Sentence-length Informed Method
- ✓Experiments and Analysis
- ✓Conclusions and Future Work

jhdu@xaut.edu.cn

Sentence-length Informed Method

MOTIVATION

✓ Problem: The rich-information sentence selection algorithm prefers to pick up short sentences;

✓ Solution: We filtered out short sentences to guarantee the effectiveness of the algorithm in previous work;

✓ Consequence: We lost many useful sentences that results in lower translation performance;

✓ Our Intuition: Introducing a penalty factor to punish the shorter sentences to reach a balance between data and performance.

jhdu@xaut.edu.cn

Translation Units-based Selection Algorithms

●Basic Algorithms

❖Geom-Phrase Algorithm

The Geom-Phrase algorithm is defined as

$$\phi(s) := [\prod_{x \in X_s^m} \frac{P(x|U)}{P(x|L)}]^{\frac{1}{|X_s^m|}} \quad (1)$$

❖ Arith-Phrase Algorithm

The Arith-Phrase algorithm is defined as

$$\phi(s) := \frac{1}{|X_s^m|} \sum_{x \in X_s^m} \frac{P(x|U)}{P(x|L)} \quad (2)$$

X_s^m	The set of possible phrases that the sentence s can offer
$P(x\|D) = \frac{Count(x)+\varepsilon}{\sum_{x \in X_D^P} Count(x)+\varepsilon}$	The probabilities of observing x in U and L respectively, ε is the smooth factor.

jhdu@xaut.edu.cn

Proposed Sentence-length Informed Algorithm

- **Modified Arith-Phrase Algorithm with a Penalty Factor**

  - the sentence length has a significant impact on the selection performance.
  - The modified Arith-Phrase algorithm which we call it "Arith-Phrase-Penalty" is as in Eq. (3)

$$\phi(s) = [\frac{1}{|X_s^m|} \sum_{x \in X_s^m} \frac{P(x|U)}{P(x|L)}] \times BP \quad (3)$$

where *BP* is the brevity penalty and defined as follows,

$$BP = \begin{cases} 1 & if\ c > r \\ e^{1-\frac{r}{c}} & if\ c \le r \end{cases} \quad (4)$$

where *r* is the average sentence length in the monolingual corpus *U*, *c* is length of the sentence to be selected.

jhdu@xaut.edu.cn

Proposed Sentence-length Informed Algorithm

ALGORITHM MECHANISM

✓ The penalty is calculated by the ratio of the current candidate sentence length ***c*** and the overall average sentence length ***r*** of the monolingual corpus;

✓ The penalty is not fixed that is *dynamically updated* at each iteration of sentence selection process.

jhdu@xaut.edu.cn

Outline

- ✓Introduction
- ✓Sentence-length Informed Method
- ✓Experiments and Analysis
- ✓Conclusions and Future Work

jhdu@xaut.edu.cn

Experiments and Analysis

Experimental Settings – SMT system

- **Language pair:** Chinese-English
- **Training set:** NIST FBIS corpus
- **Devset:** NIST 2006 current set
- **Testset:** NIST 2005, 2008 sets

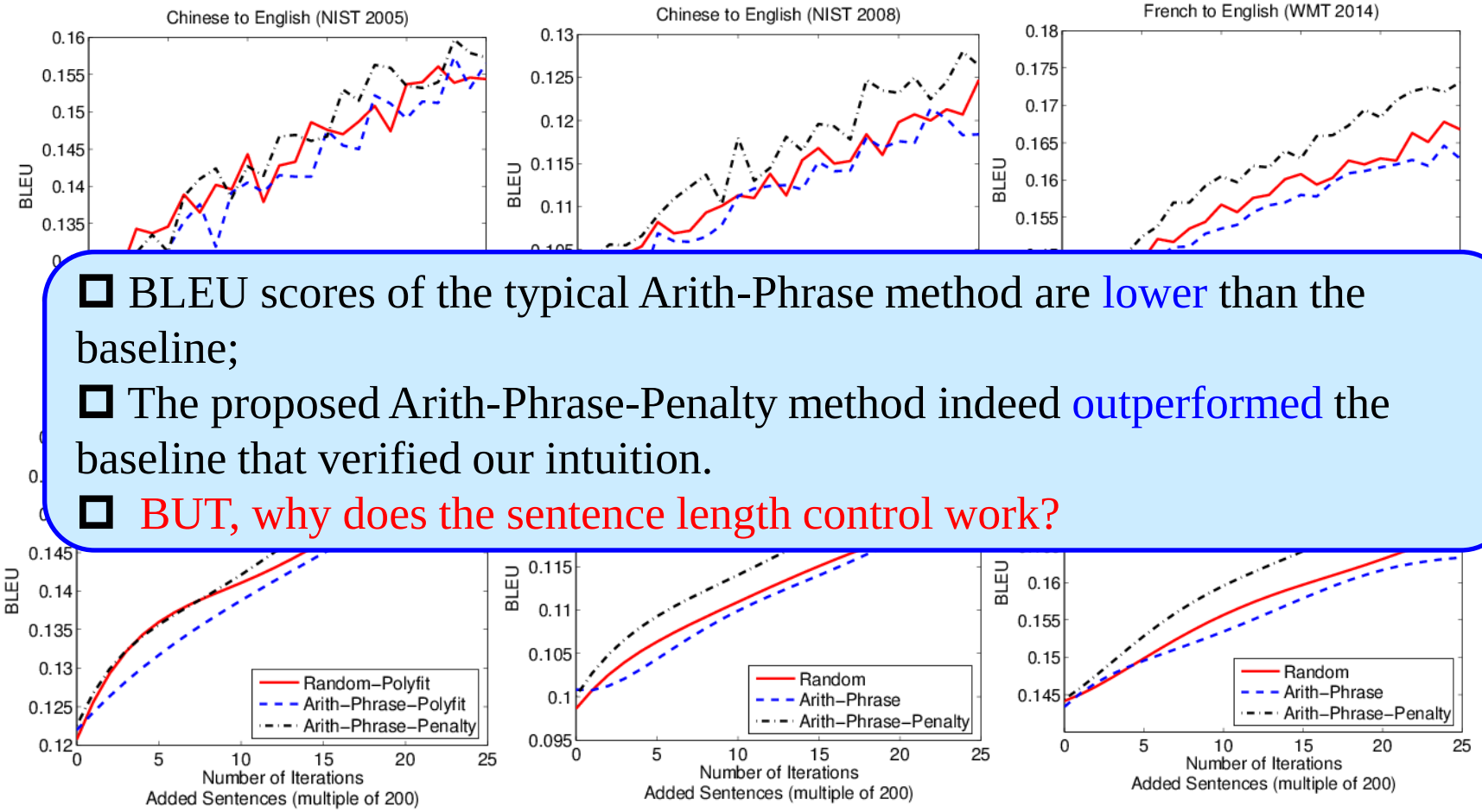
- **Language pair:** French-English
- **Training set:** WMT News Commentary corpus
- **Devset:** WMT Newsteset 2013
- **Testset:** WMT Newsteset 2014

✓ The initial parallel data contains **5k** pairs and the monolingual data includes **20k** sentences;
✓ Baseline: Random selection

jhdu@xaut.edu.cn

Experiments and Analysis

- **Experimental results**



- BLEU scores of the typical Arith-Phrase method are lower than the baseline;
- The proposed Arith-Phrase-Penalty method indeed outperformed the baseline that verified our intuition.
- BUT, why does the sentence length control work?

Fig.2.Experimental results of Arith-Phrase, Sentence-length Informed (Arith-Phrase-Penalty) methods compared to the baseline (Random)

jhdu@xaut.edu.cn

Experiments and Analysis

Data Statistics and Observations:

① Sentences by Arith-Phrase method are shorter than those selected by Random method and Arith-Phrase-Penalty.

➢ **Average Sentence Length**

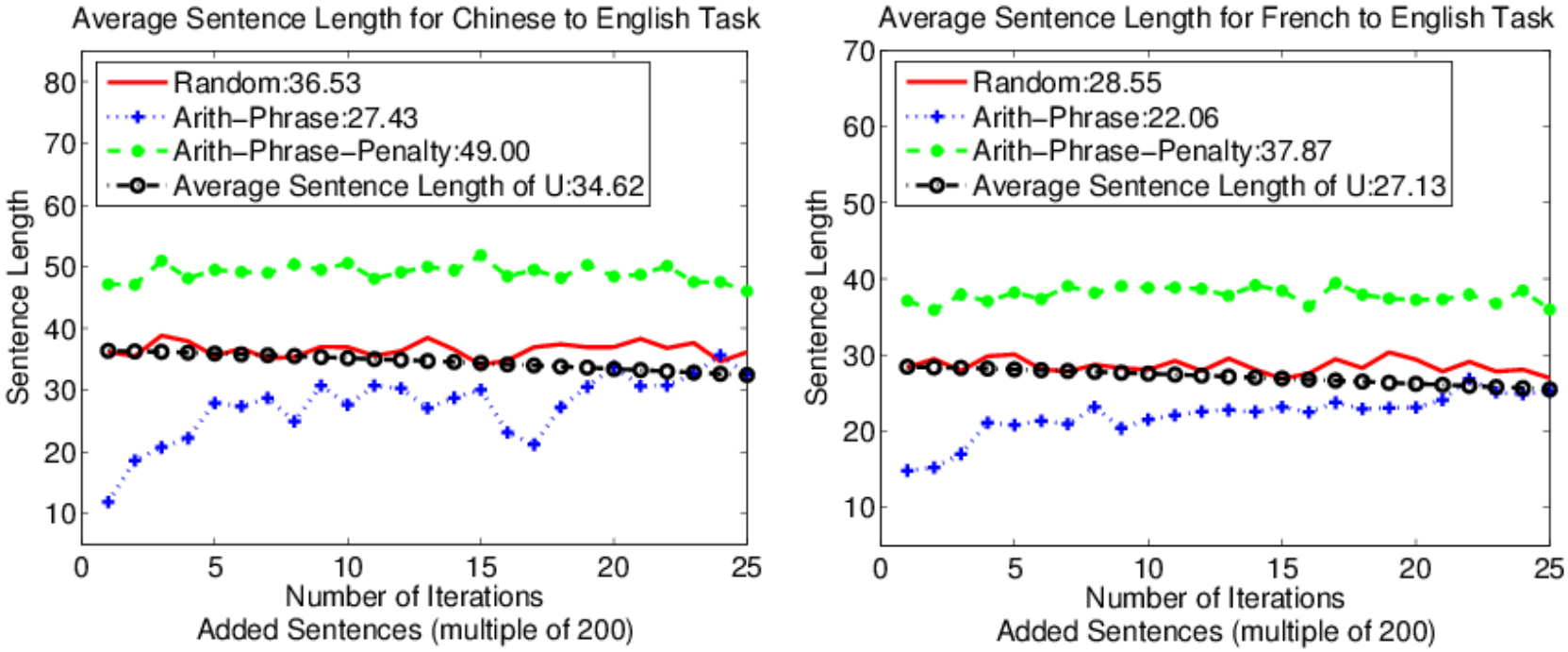


Fig. 3. Statistics of sentence length for different selection methods

Experiments and Analysis

Data Statistics and Observations:

② Testset coverage of the Arith-Phrase is higher than that of the Random method, but lower than Arith-Phrase-Penalty.

➢ **Coverage of Testsets by Parallel Data**

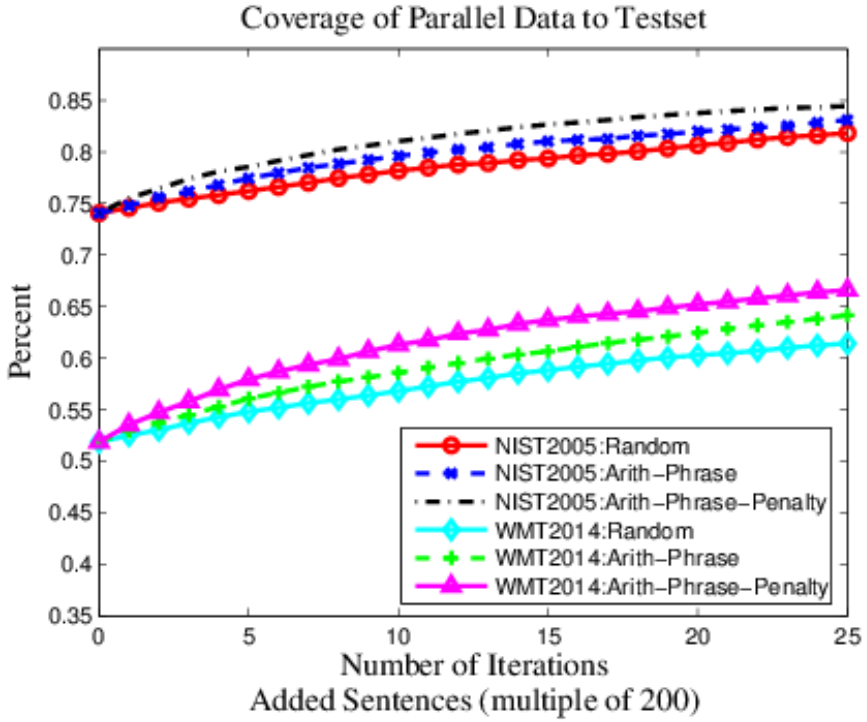


Fig.4.Comparison of coverage rates of parallel data to different test sets at each iteration

jhdu@xaut.edu.cn

Experiments and Analysis

●Analysis

■A pair of contradiction: **exploration** and **exploitation**

➢ selecting sentences to discover new phrases **vs** estimating accurately the phrase translation probabilities ➔ a tradeoff

■**Inference:**

➢ the more new words a sentence has in terms of the parallel corpus, the more informative the sentence is, but a lower word alignment accuracy to the added parallel data

➢ while the more existing words a sentence has to the parallel data, the more accurate the phrase probability is estimated, but a lower coverage to the test set

✓ the relationship between "existing words" and "new words" is crucial!

✓ the variation between the frequencies of existing words and new words might provide a reasonable explanation.

jhdu@xaut.edu.cn

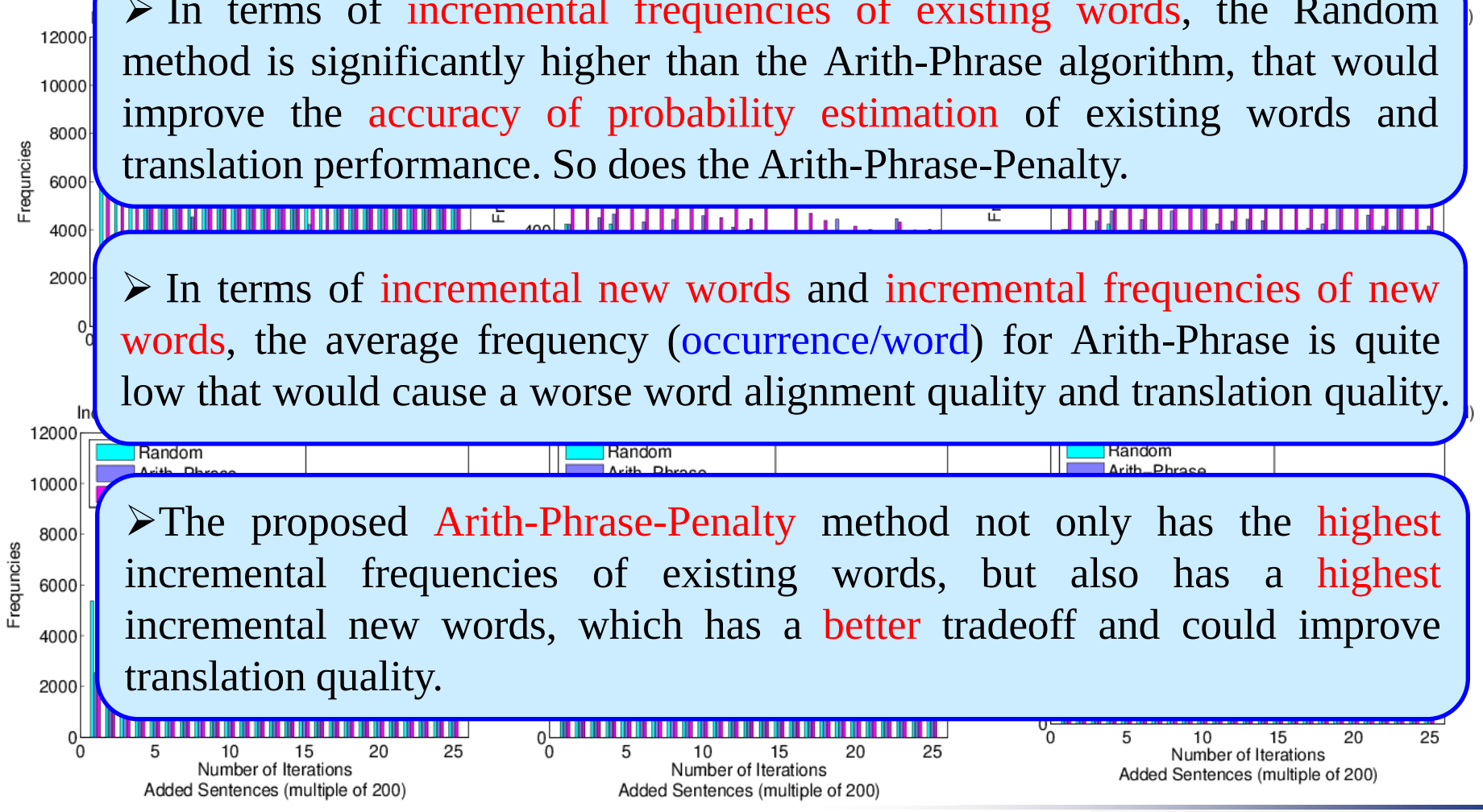
Experiments and Analysis

➢ **Statistics of Existing Words and New Words**

➢ In terms of incremental frequencies of existing words, the Random method is significantly higher than the Arith-Phrase algorithm, that would improve the accuracy of probability estimation of existing words and translation performance. So does the Arith-Phrase-Penalty.

➢ In terms of incremental new words and incremental frequencies of new words, the average frequency (occurrence/word) for Arith-Phrase is quite low that would cause a worse word alignment quality and translation quality.

➢The proposed Arith-Phrase-Penalty method not only has the highest incremental frequencies of existing words, but also has a highest incremental new words, which has a better tradeoff and could improve translation quality.



jhdu@xaut.edu.cn

Outline

- ✓Introduction
- ✓Sentence-length Informed Method
- ✓Experiments and Analysis
- ✓Conclusions and Future Work

jhdu@xaut.edu.cn

Conclusions and Future Work

➢Conclusions

① Based on the negative experimental results on Arith-Phrase method, we found that the sentence length is an important factor to affect the system performance when the length of sentences in the monolingual corpus varies in a wide range.

② A simple but effective method – sentence length informed Arith-Phrase – is proposed to penalize shorter sentences to reach a better tradeoff.

③ Experimental results demonstrate that the proposed method significantly outperforms the typical Arith-Phrase and Random method.

jhdu@xaut.edu.cn

Conclusions and Future Work

Q

The increased the sentence length of the proposed method will also increase the cost of human translation in AL framework, which might not be acceptable in practical use.

- Future Work

1

A user study to verify how much human effort would increase, and how much it would be accepted with the length increase.

2

Better algorithms to make a better tradeoff and to select more informative sentences.

jhdu@xaut.edu.cn

陕西省复杂系统控制与
智能信息处理重点实验室
Laboratory for Complex System Control
and Intelligent Information Processing

THANKS FOR YOUR ATTENTION!

jhdu@xaut.edu.cn

西安理工大学
XI'AN UNIVERSITY OF TECHNOLOGY