

# MSAMSum: Towards Benchmarking Multi-lingual Dialogue Summarization

**Xiachong Feng** , Xiaocheng Feng , Bing Qin

Research Center for Social Computing and Information Retrieval  
Harbin Institute of Technology

# Dialogue summarization

- Distill the most important information from a dialogue (**content selection**)
  - Meeting, Email, Chit-chat, Customer Service, Debate .....
- Convert them into a short textual passage (**surface realization**)



## Current works

- Various dialogue summarization datasets are proposed to meet different needs and scenarios.
- Current works mainly focus on **English** dialogue summarization, leaving other languages less well explored.

Name	Domain	Language
ICSI [Janin <i>et al.</i> , 2003]	Meeting	English
AMI [Carletta <i>et al.</i> , 2005]		English
QMSum [Zhong <i>et al.</i> , 2021]		English
SAMSum [Gliwa <i>et al.</i> , 2019]	Chat	English
GupShup [Mehnaz <i>et al.</i> , 2021]		Code-Mix
CSDS [Lin <i>et al.</i> , 2021]	Customer Service	Chinese
TODSum [Zhao <i>et al.</i> , 2021]		English
TWEETSUMM [Feigenblat <i>et al.</i> , 2021]		English
CRD3 [Rameshkumar and Bailey, 2020]	TV Show	English
[Song <i>et al.</i> , 2020]	Medical	Chinese
SumTitles [Malykh <i>et al.</i> , 2020]	Movie	English
MEDIASUM [Zhu <i>et al.</i> , 2021]	Interview	English
DIALOGSUM [Chen <i>et al.</i> , 2021]	Spoken	English
EMAILSUM [Zhang <i>et al.</i> , 2021a]	Email	English
ForumSum [Khalman <i>et al.</i> , 2021]	Forum	English
ConvoSumm [Fabbri <i>et al.</i> , 2021]	Mix	English

Table 1: Major datasets for dialogue summarization.

# How about a multi-lingual dialogue summarizer?



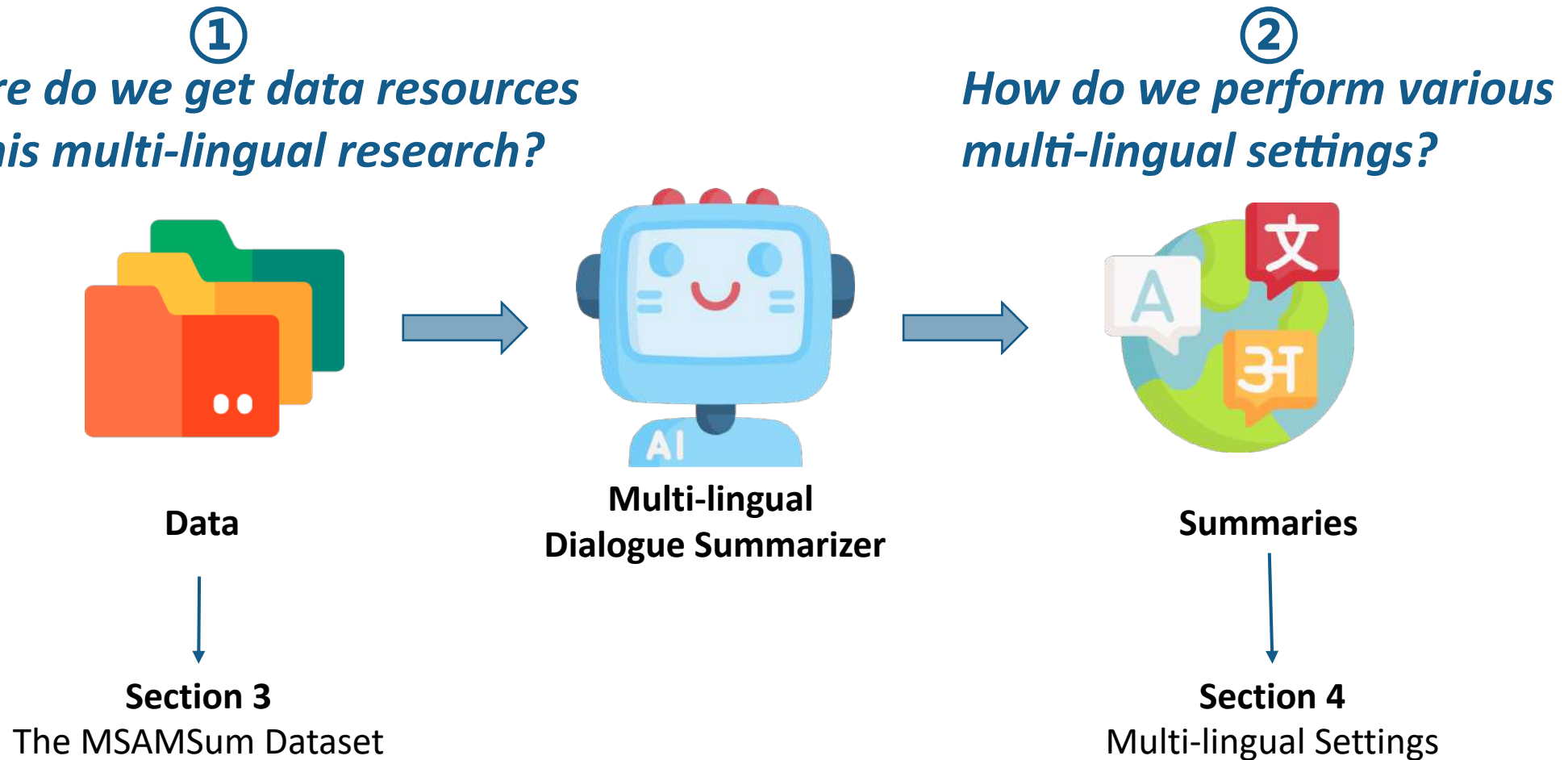
## Summary Service Provider

- |                   |                   |
|-------------------|-------------------|
| Ⓔ English Service | Ⓔ Russian Service |
| Ⓔ Chinese Service | Ⓔ French Service  |

With the help of **instantaneous translation systems**, a dialogue involving multinational participants becomes more and more common and frequent.

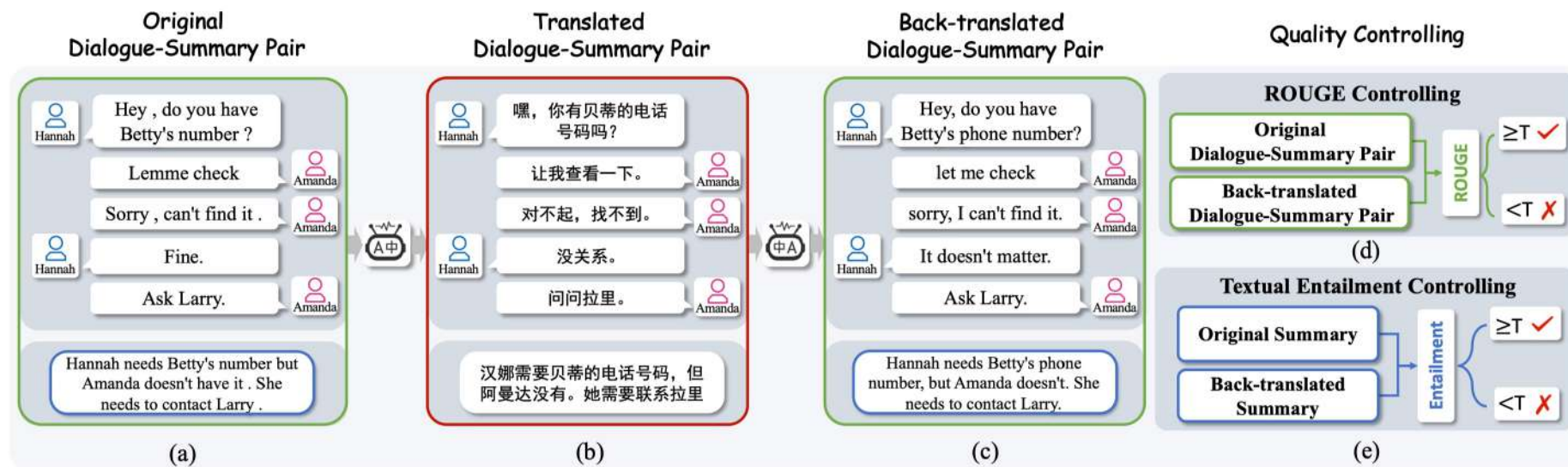
★ It is valuable to provide them with dialogue summaries **in a preferred language**.

# What should we do?



# MSAMSum Dataset

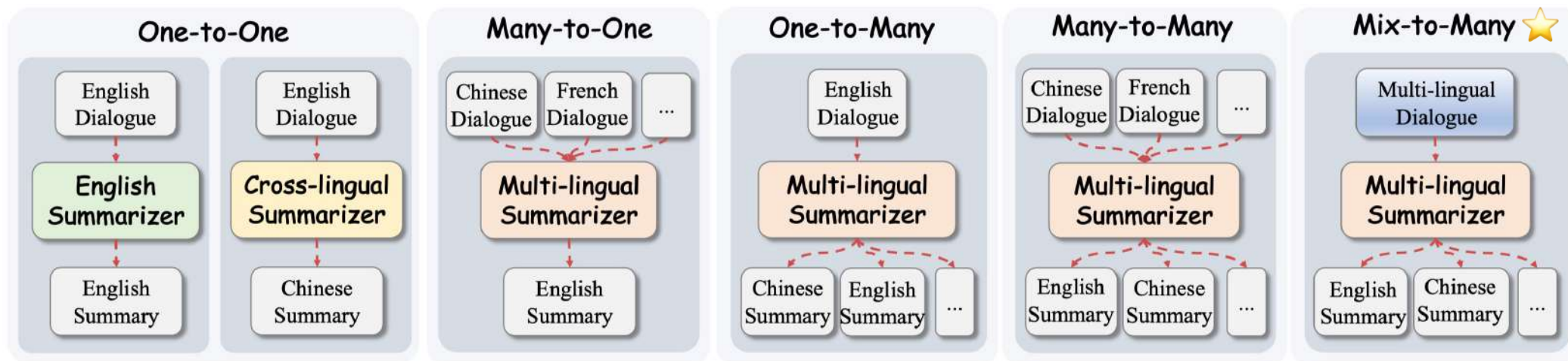
- **SAMSum**: human-labeled large-scale dataset; covers a wide range of domains.
- **Six official languages of the United Nations**: Arabic, Chinese, English, French, Russian and Spanish
- **Quality Controlling**:
  - *Round-trip Translation*: coarse-grained surface-level quality
  - *Textual Entailment*: fine-grained factual consistency





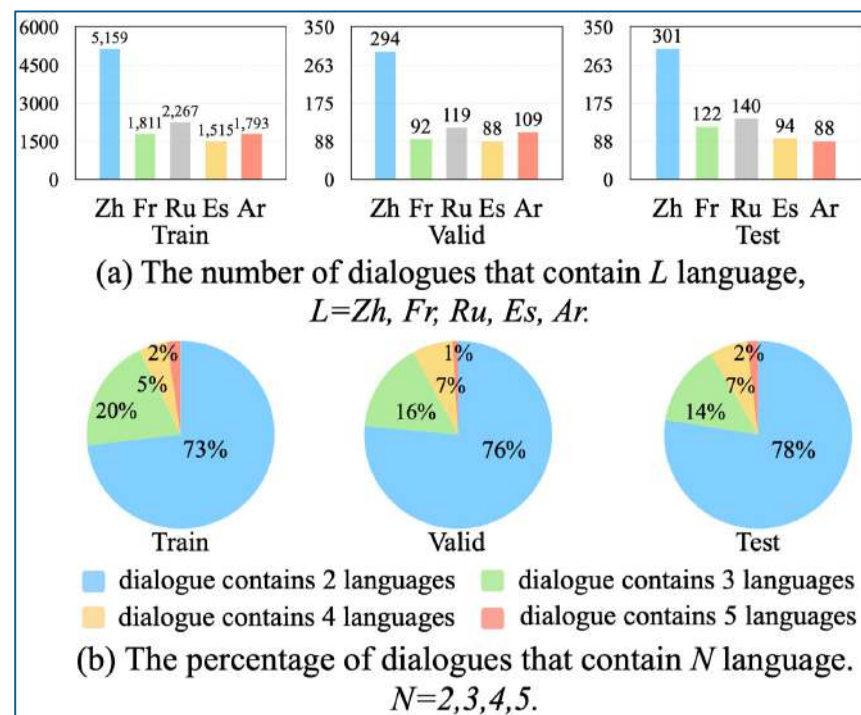
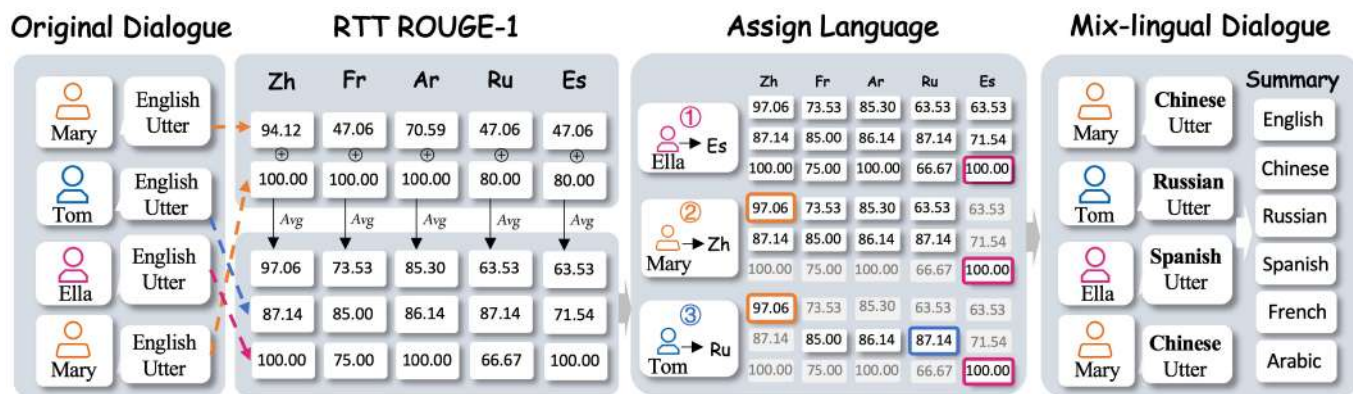
# Multi-lingual Settings

- **One-To-One** is the basic setting.
- **Many-To-One** model encodes  $N$  languages and decodes to English.
- **One-To-Many** model encodes English and decodes into  $N$  languages.
- **Many-To-Many** model encodes and decodes  $N$  languages.
- ★ **Mix-To-Many**, where the model takes a mix-lingual dialogue (utterances in a dialogue belongs to different languages) as input and outputs summaries in different languages.



# How to construct Mix-To-Many Setting?

- We first group utterances for the same participant and get the averaged round-trip translation ROUGE-1 score for each language.
- Then, we adopt a greedy search strategy to assign each participant a language.
- Finally, we can get the mix-lingual dialogue associated with summaries in different languages.

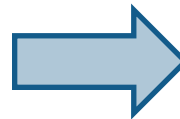


Statistics for mix-lingual dialogues



# Case

English Dialogue	
<b>Elliot</b>	: I can't talk rn , I'm rly busy.
<b>Elliot</b>	: Can I call u back in about 2 hours?
<b>Jordan</b>	: Not really , I'm going to a funeral.
<b>Jordan</b>	: I'll call you tonight , ok?
<b>Elliot</b>	: Sure
<b>Elliot</b>	: Whose funeral is it?
<b>Jordan</b>	: My colleague's , <b>Brad</b> .
<b>Jordan</b>	: I told you about him , he had a <b>liver cancer</b> .
<b>Elliot</b>	: I'm so sorry man , I hope u're ok.
<b>Elliot</b>	: I'll call u at <b>8 pm</b> .



Generated Summaries ( <i>One-to-many</i> )	
English	<b>Elliot</b> can't talk because he's busy. <b>Jordan</b> is going to a funeral for his colleague, <b>Brad</b> , who had a <b>liver cancer</b> . <b>Elliot</b> will call him at <b>8 pm</b> . [71.19-42.11-50.85]
Chinese	乔丹要去参加他的同事 <b>布拉德</b> 的葬礼。他得了 <b>肝癌</b> 。 <b>埃利奥特</b> 将在 <b>晚上8点</b> 给乔丹打电话。 [66.67-40.00-35.09]
Russian	<b>Джордан</b> собирается на похороны своего коллеги <b>Брэда</b> , у него <b>рак печени</b> . <b>Элиот</b> позвонит Джордана в <b>20: 00</b> . [58.38-30.00-38.10]
French	<b>Elliot</b> ne peut pas parler parce qu'il est occupé. <b>Jordan</b> va au funeral de son collègue, <b>Brad</b> , qui a un <b>cancer du foie</b> . Il appellera <b>Elliot</b> à <b>20 h</b> . [68.97-42.86-55.17]
Arabic	<b>جوردن</b> هو الذهاب إلى جنازة زميلها <b>براد</b> لديه <b>سرطان الكبد</b> . <b>إيليت</b> سوف ندعو له في الساعة <b>الثامنة مساء</b> . [57.78-27.91-31.11]
Spanish	<b>Elliot</b> no puede hablar porque está ocupado. <b>Jordan</b> va a un funeral de su colega, <b>Brad</b> , que tuvo un <b>cáncer de hepática</b> . <b>Eliot</b> llamará a Jordan a las <b>8 p.m</b> . [60.71-29.63-39.29]

 See more details in the poster session~