Multi-Source Neural Machine Translation with Data Augmentation

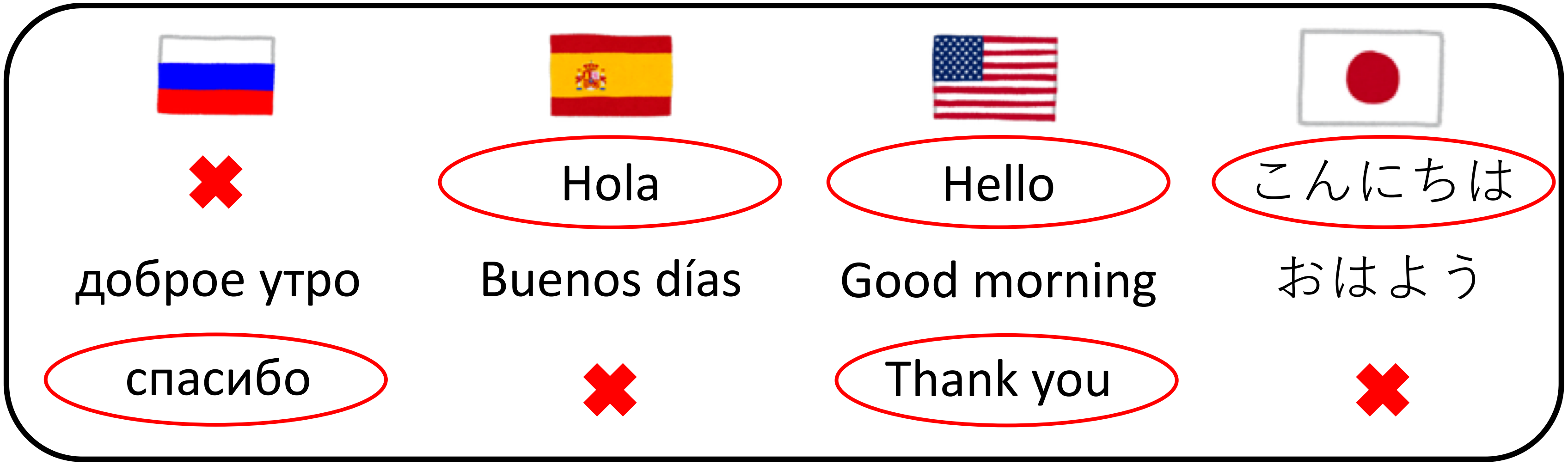
Yuta Nishimura[1], Katsuhito Sudoh[1], Graham Neubig[2,1], Satoshi Nakamura[1]

[1]Nara Institute of Science and Technology

[2]Carnegie Mellon University

Overview of this research (1/2)

Multi-lingual corpora usually have missing translations



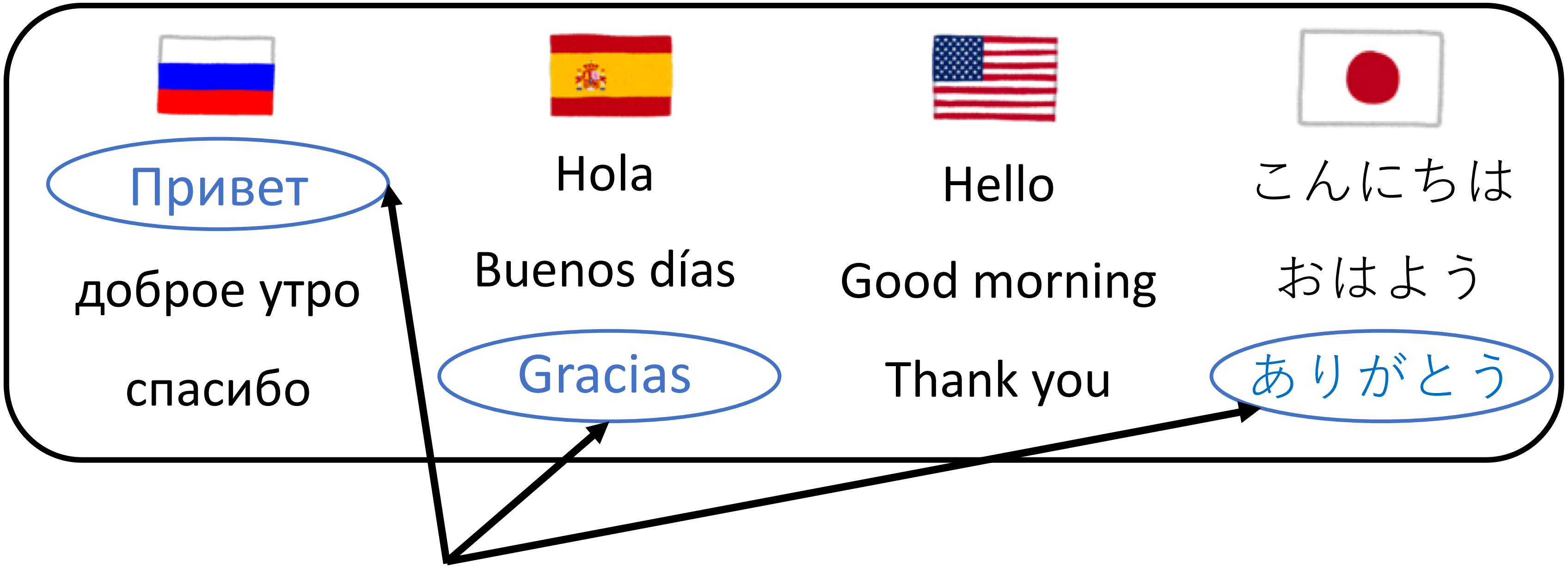
🇷🇺	🇪🇸	🇺🇸	🇯🇵
✖	Hola	Hello	こんにちは
доброе утро	Buenos días	Good morning	おはよう
спасибо	✖	Thank you	✖

In multi-source machine translation,
we cannot use the translation surrounded by a red circle

We would like to use all available translations

Overview of this research (2/2)

We would like to use all available translations



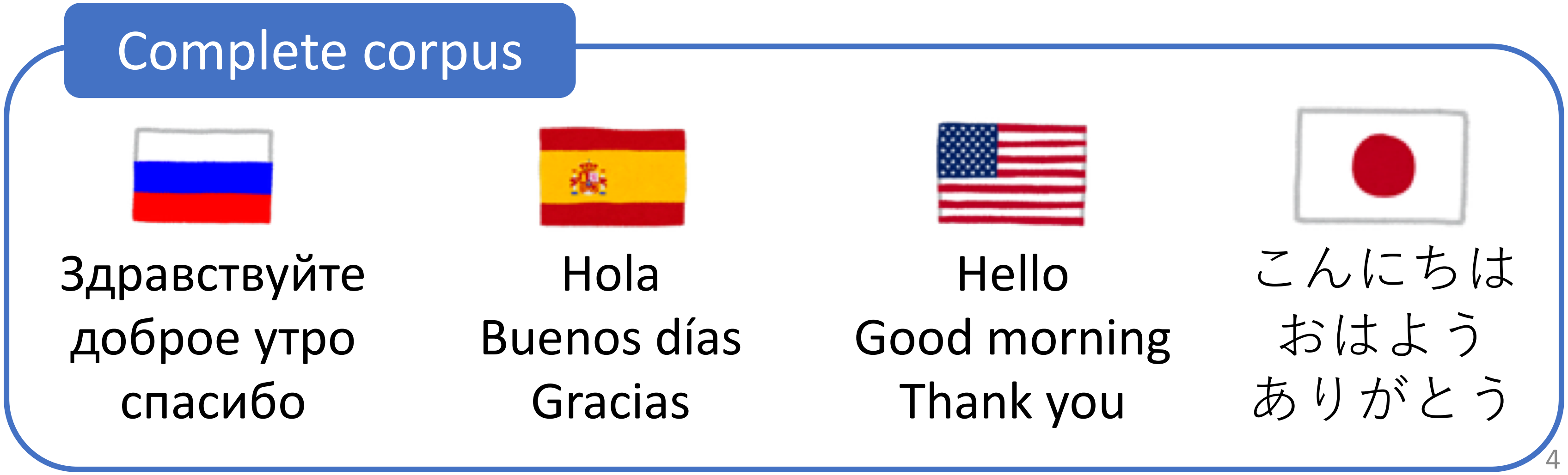
Привет	Hola	Hello	こんにちは
доброе утро	Buenos días	Good morning	おはよう
спасибо	Gracias	Thank you	ありがとう

- We augment with pseudo-translations using multi-source NMT
- Our proposed method achieved good result

Multi-lingual Corpus

- There are many corpora which have multiple languages
  - Video captions for talks or movies [Cettolo et al., 2012; Tiedemann, 2009]
  - Europarl [Kohen, 2005], UN [Ziemski et al., 2016]

These corpora have good, manually curated translations in a number of languages



Complete corpus

Здравствуйте	Hola	Hello	こんにちは
доброе утро	Buenos días	Good morning	おはよう
спасибо	Gracias	Thank you	ありがとう

Multi-lingual corpus with missing data

It is unusual that sentences of all languages exist

(such as subtitles for TED Talks)

⬇

Goal

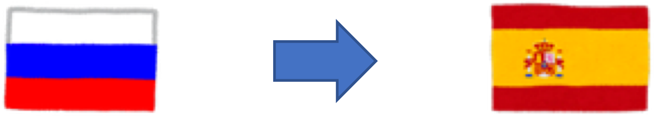
Generating good translations in the remaining languages for which do not yet have translations in a multilingual corpus

Incomplete corpus

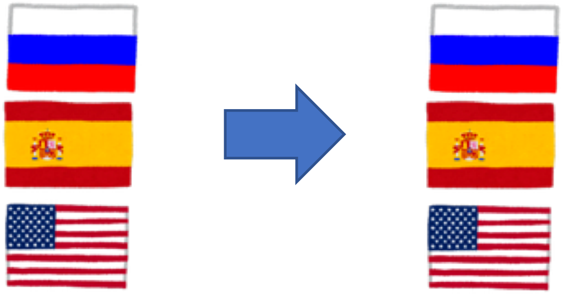
✖	Hola	Hello	こんにちは
доброе утро	Buenos días	Good morning	おはよう
спасибо	✖	Thank you	✖

Neural Machine Translation (NMT)

One-to-one NMT



Multi-lingual NMT



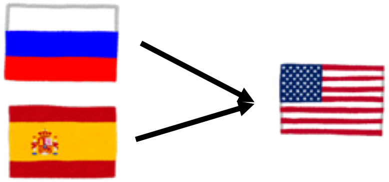
☺ Better!

We use multi-lingual NMT to generate translations

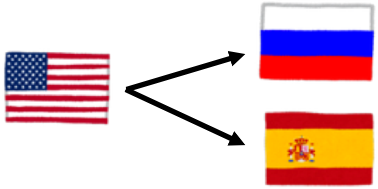
But there are some types of Multi-lingual NMT

Multi-lingual NMT

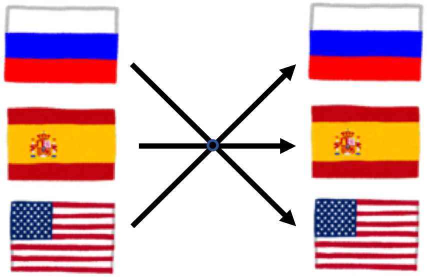
- Multi-Source, One-Target
  [Zoph and Knight, 2016; Garmash and Monz, 2016]



- One-Source, Multi-Target
  [Firat et al., 2016]



- Multi-Source, Multi-Target
  [Johonson et al., 2017; He et al., 2016]

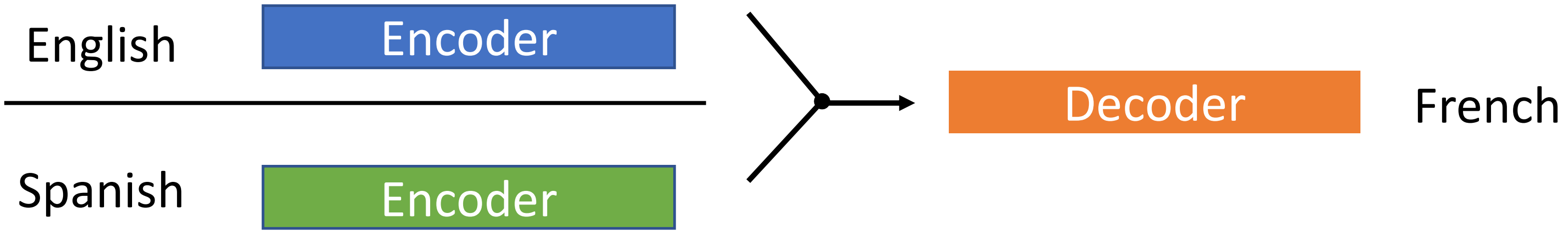


We'd like to improve NMT by the help of
the other curated translations on the source side at the test time

We focused on Multi-Source and One-Target

Multi-Source NMT | Multi-Encoder NMT

- Multi-Encoder NMT [Zoph and Knight, 2016]
  - Multiple Encoder and one Decoder
  - Multiple sentences are each encoded separately, then all referenced during decoding process



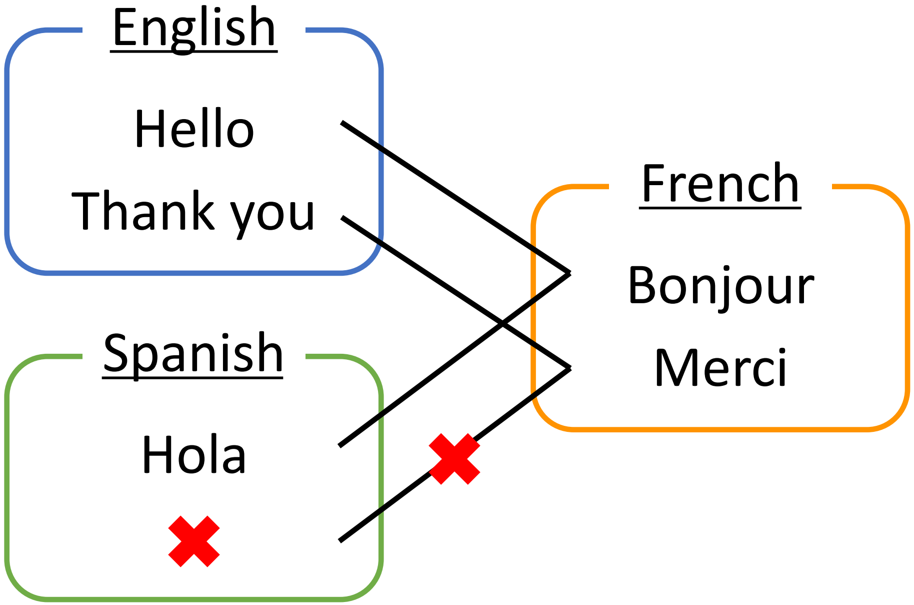
The disadvantage of Multi-Source NMT

Multi-Source NMT assumes we have data in all of languages

↓

we cannot use the translation

if some source translations are missing

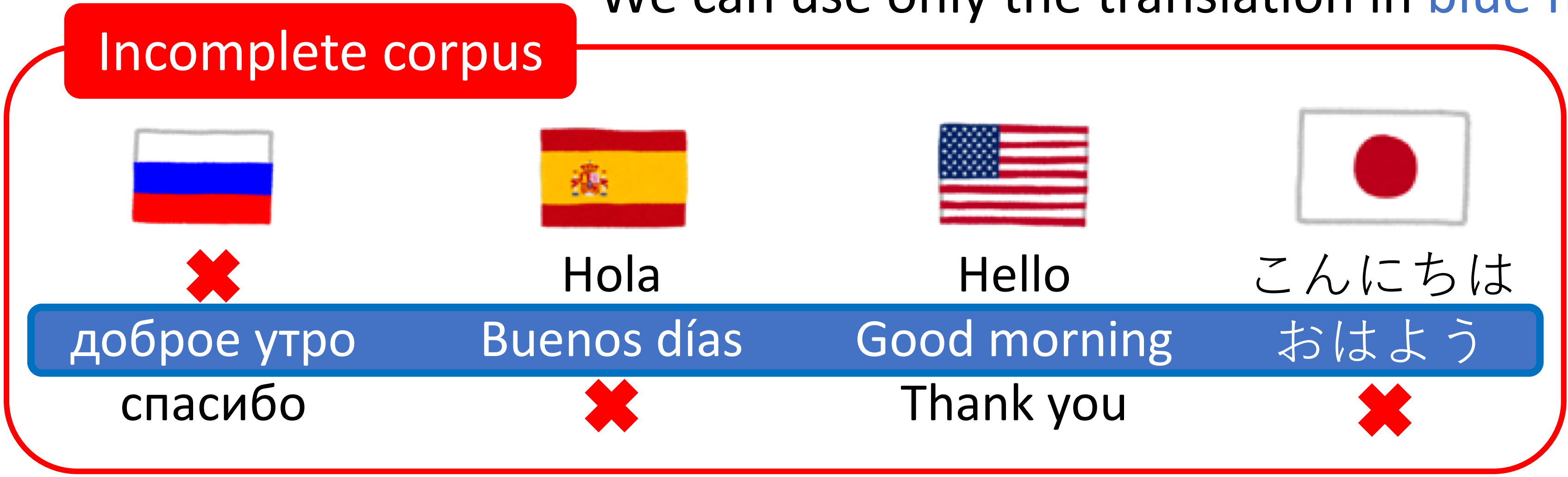


We cannot use the translation
“Thank you” and “Merci”
in multi-source NMT

About our research

We would like to use all available translations even if the corpus has missing data

We can use only the translation in blue frame



Incomplete corpus

✖	Hola	Hello	こんにちは
доброе утро	Buenos días	Good morning	おはよう
спасибо	✖	Thank you	✖

Our research is **the first study** on how to handle incomplete corpora

Our Previous Work

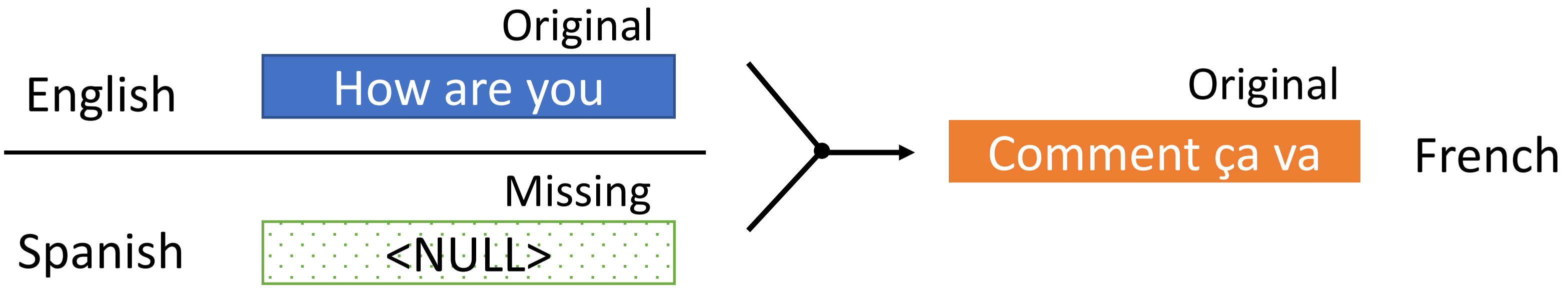
Problem

Multi-Source NMT assumes that we do not have any missing data

Proposed

Replacing each missing input sentence with a special symbol <NULL>

[Nishimura et al., 2018]



This method achieved higher translation accuracy

Our Previous Work’s Problem

In case, the corpus has many missing data

The model will be trained on corpora with a large number of NULL symbols

Problem

The source condition will be much different between train time and test time

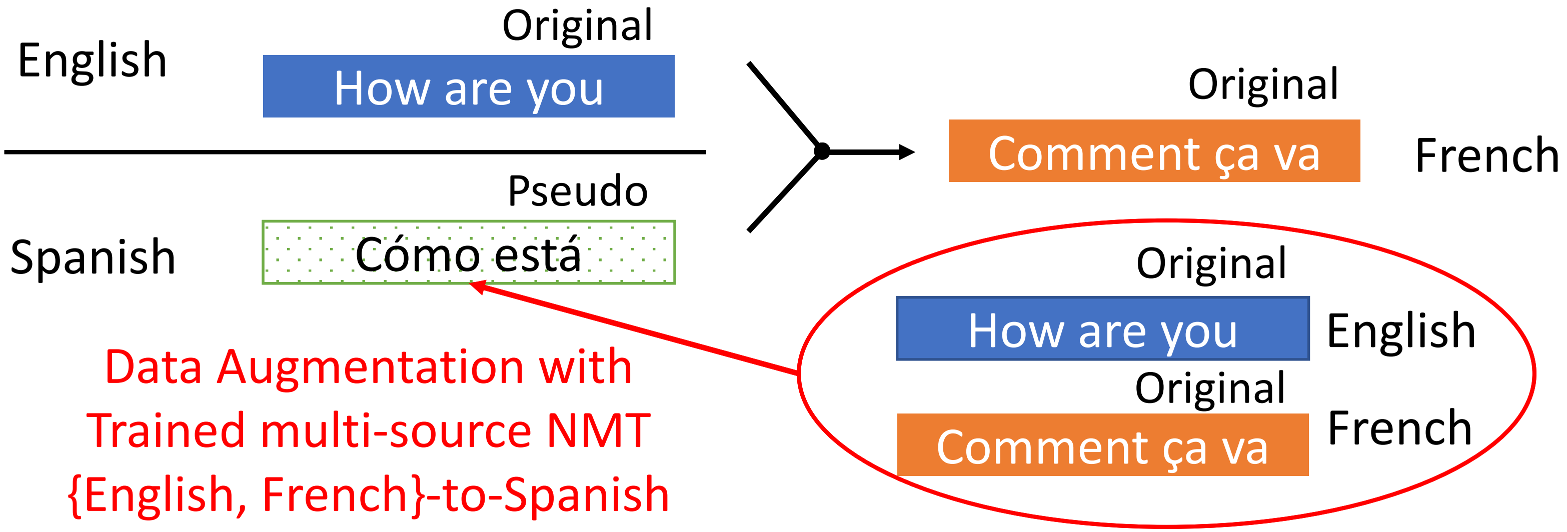
Proposed Method | Overview

Problem

The source condition is very different between train and test

Proposed

Using a pseudo-corpus that fills missing data with multi-source NMT outputs

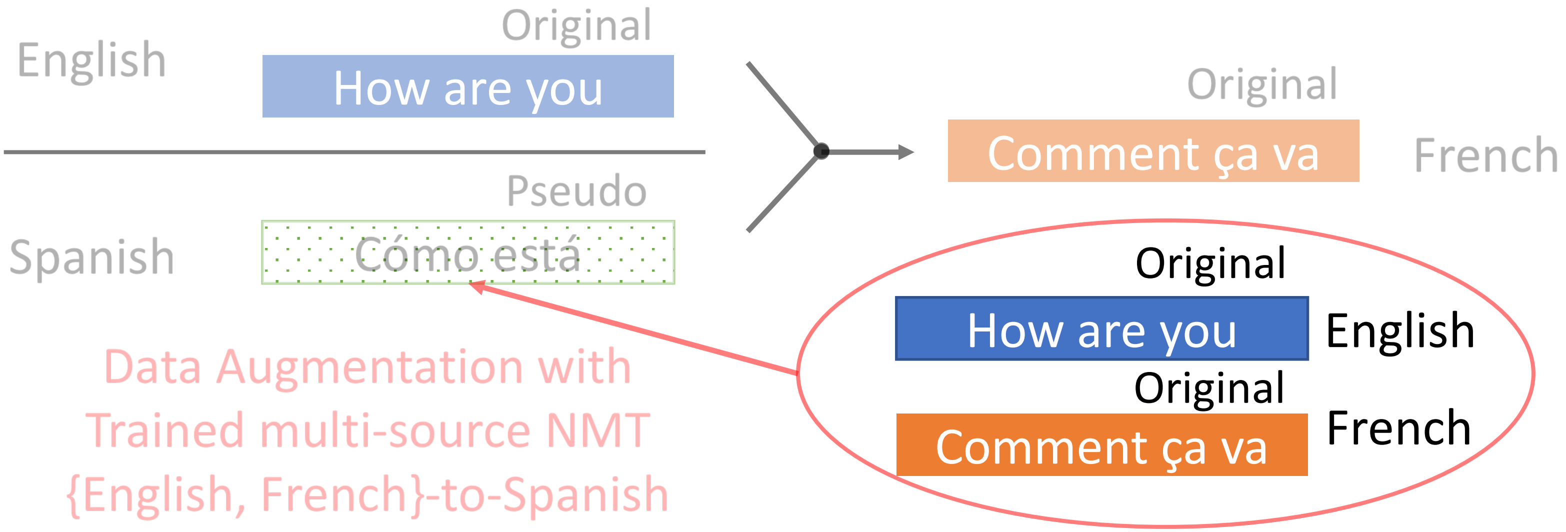


Data Augmentation with Trained multi-source NMT {English, French}-to-Spanish

Proposed Method | 1st step

Final Goal : Get French Translation

- Train a multi-encoder NMT model (Source: English and French, Target: Spanish)
- If there is a missing input, we replace a missing input sentence with a special symbol <NULL>

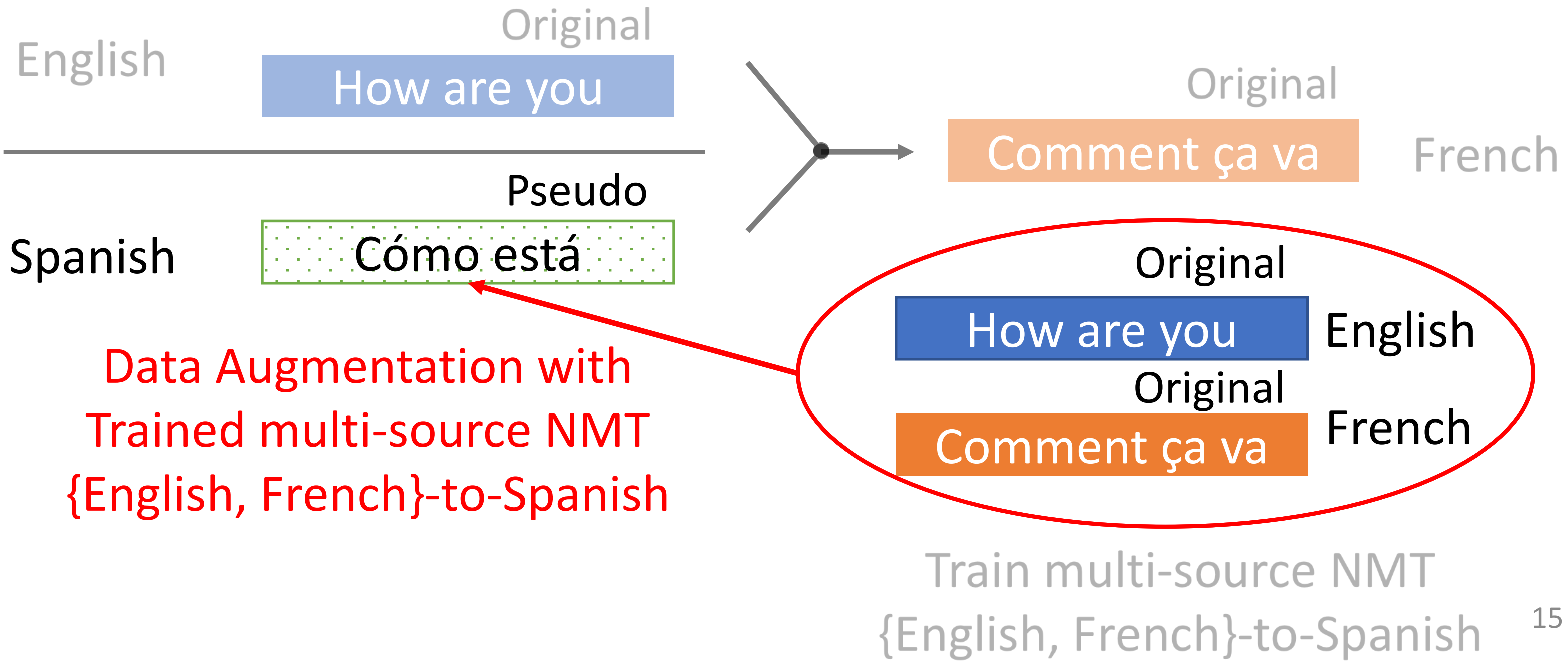


Data Augmentation with Trained multi-source NMT {English, French}-to-Spanish

Train multi-source NMT {English, French}-to-Spanish

Proposed Method | 2nd step

- Create Spanish pseudo-translations using multi-encoder NMT which was trained on the 1st step
- We conducted three types of augmentation

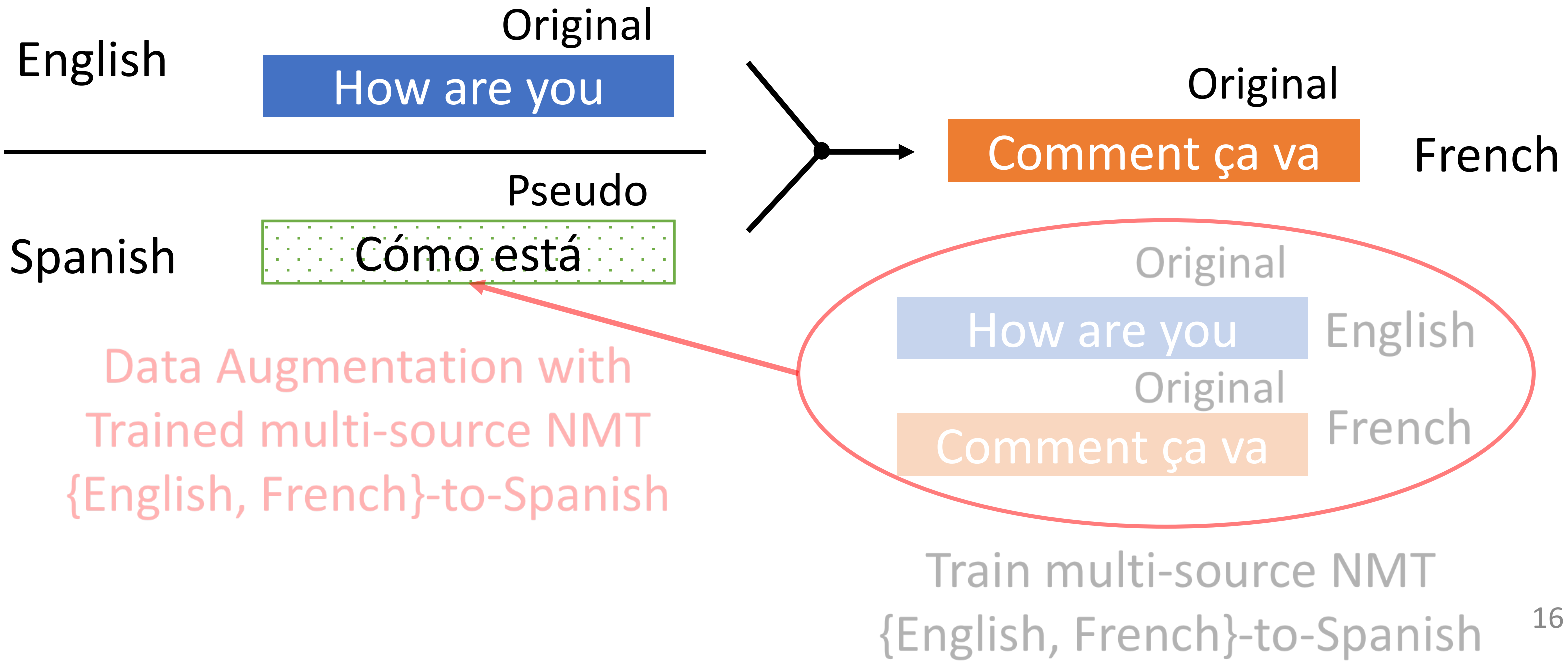


Data Augmentation with
Trained multi-source NMT
{English, French}-to-Spanish

Train multi-source NMT
{English, French}-to-Spanish

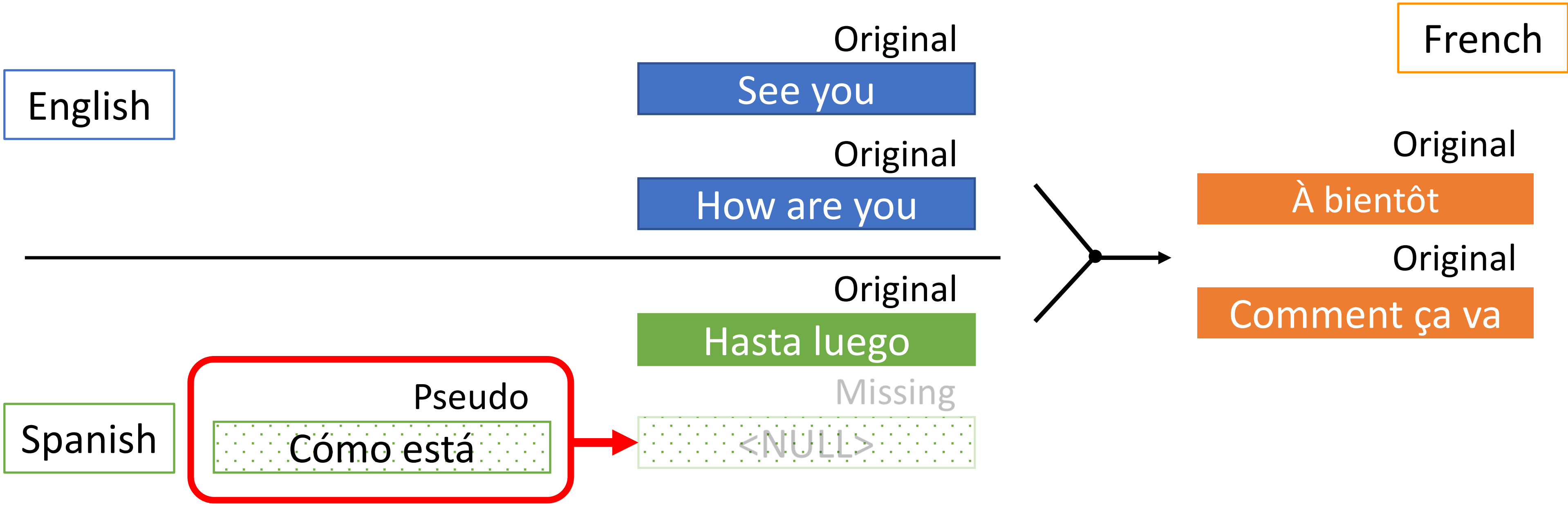
Proposed Method | 3rd step

- Train a multi-encoder NMT model (Source: English and Spanish, Target: French)
- Spanish translations have pseudo-translations



Three types of augmentation (1) : “fill-in”

- Where only missing parts in the corpus are filled up with pseudo-translations



Three types of augmentation
The reason of making three types

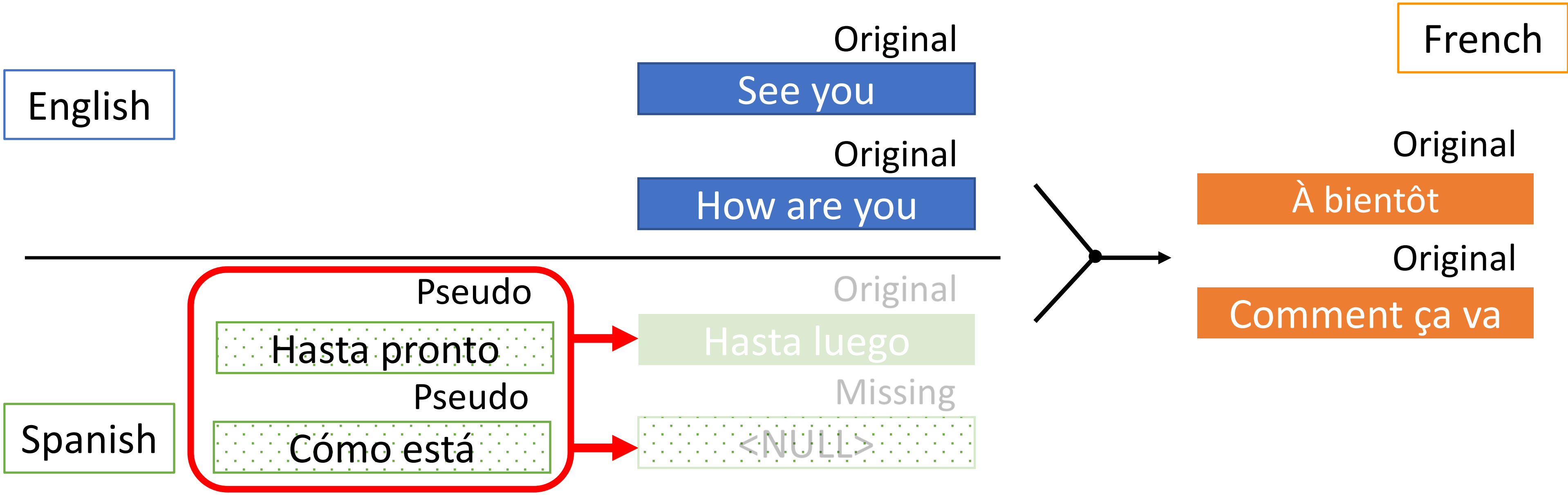
- Translations of TED talks are unreliable
  - Translations are created from many independent volunteers

The effectiveness of applying back-translation for an unreliable part of a provided corpus [Morishita et.al. 2017]

We proposed the methods not to use unreliable original translations

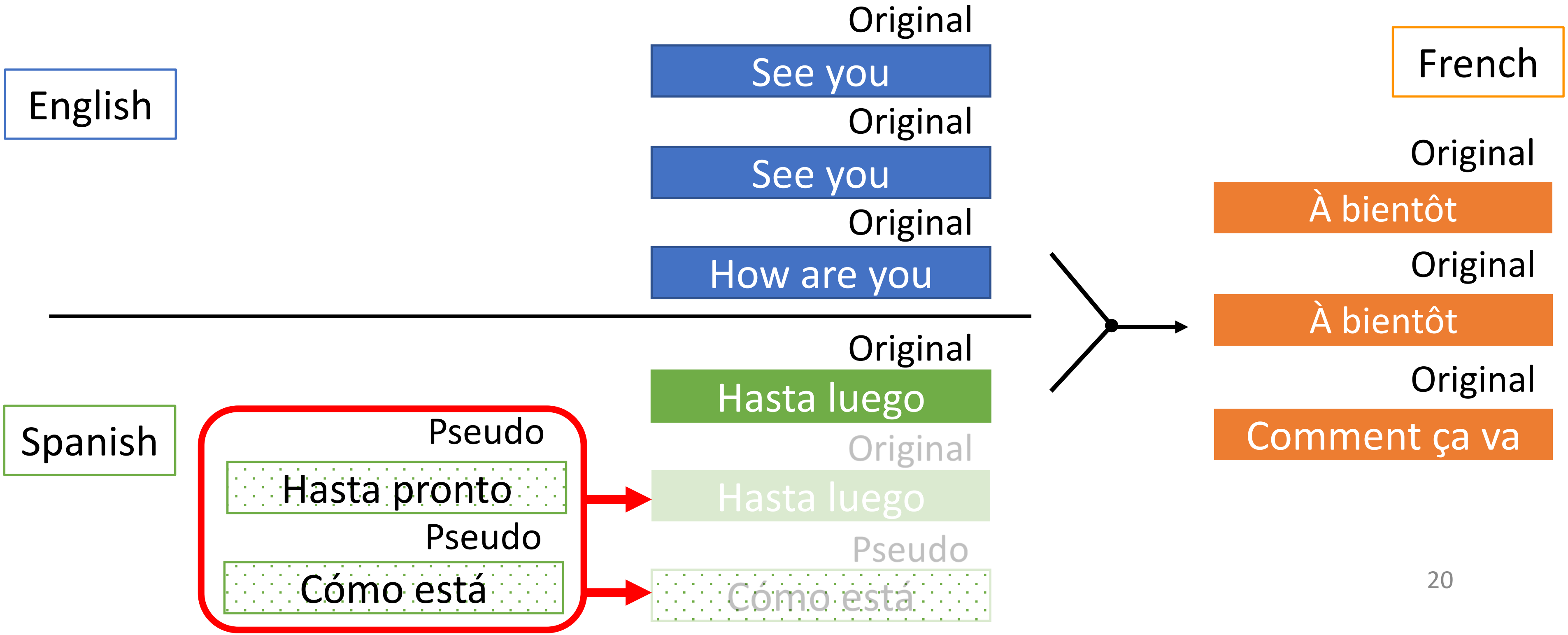
Three types of augmentation (2) : “fill-in and replace”

- Augment the missing part and replace original translations with pseudo-translations
- The motivation is not to use unreliable translation



Three types of augmentation (3) : "fill-in and add"

- Augment the missing part and added pseudo-translations from original translations
  - The motivation : prevent noise of the complete replacement of the 2nd method



Experiment | Data

- Corpus
  - A collection of transcriptions of TED Talks
- Language Pair
  - English (en), Croatian (hr), Serbian (sr)
  - English (en), Slovak (sk), Czech (cs)
  - English (en), Vietnamese (vi), Indonesian (id)

Pair	Trg	train	missing
en-hr/sr	hr	118949	35564 (29.9%)
	sr	133558	50203 (37.6%)
en-sk/cs	sk	100600	58602 (57.7%)
	cs	59918	17380 (29.0%)
en-vi/id	vi	160984	87816 (54.5%)
	id	82592	9424 (11.4%)

- train
  - the number of available training sentences
- missing
  - the number and the fraction of missing sentences in comparison with English ones

Experiment | Data

- Corpus
  - A collection of transcriptions of TED Talks
- Language Pair
  - English (en), Croatian (hr), Serbian (sr)
  - English (en), Slovak (sk), Czech (cs)
  - English (en), Vietnamese (vi), Indonesian (id)

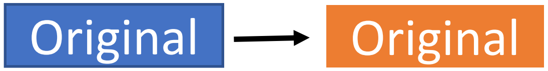
Pair	Trg	train	missing
en-hr/sr	hr	118949	35564 (29.9%)
	sr	133558	50203 (37.6%)
en-sk/cs	sk	100600	58602 (57.7%)
	cs	59918	17380 (29.0%)
en-vi/id	vi	160984	87816 (54.5%)
	id	82592	9424 (11.4%)

- train
  - the number of available training sentences
- missing
  - the number and the fraction of missing sentences in comparison with English ones

Experiment | Baseline Methods

One-to-one NMT

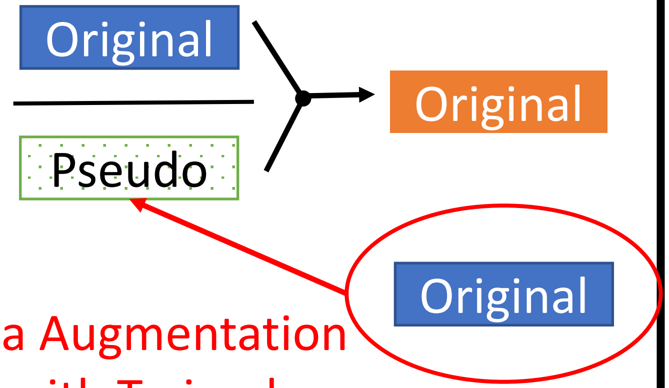
Standard NMT model from one source language to another target language



[Luong et al., 2015]

Multi-encoder NMT with back-translation

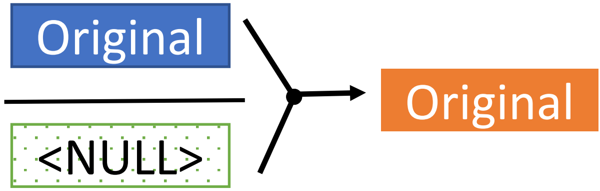
A multi encoder NMT system using pseudo-translation from English-to-X NMT



Data Augmentation with Trained one-to-one NMT English-to-X

Multi-encoder NMT with <NULL>

A multi-encoder NMT system using a special symbol <NULL>

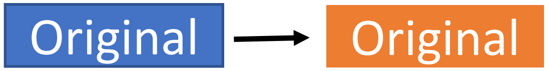


[Nishimura et al., 2018]

Baseline | One-to-one NMT

One-to-one NMT

Standard NMT model from one source language to another target language



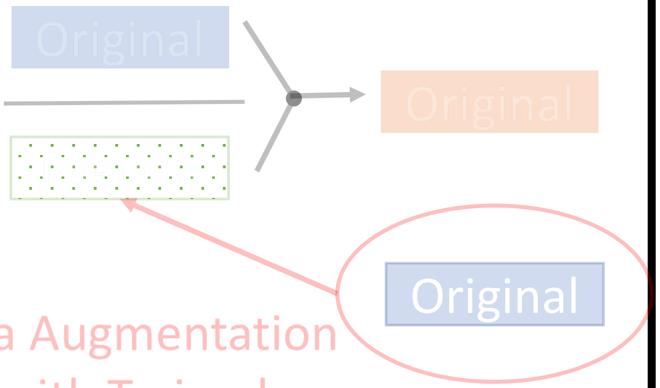
Original

Original

[Luong et al., 2015]

Multi-encoder NMT with back-translation

A multi encoder NMT system using pseudo-translation from English-to-X NMT



Original

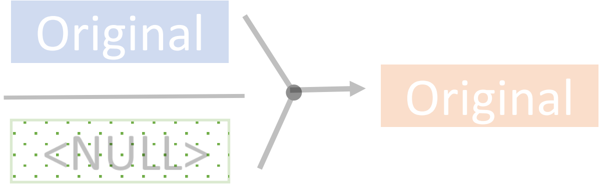
Original

Original

Data Augmentation with Trained one-to-one NMT English-to-X

Multi-encoder NMT with <NULL>

A multi-encoder NMT system using a special symbol <NULL>



Original

<NULL>

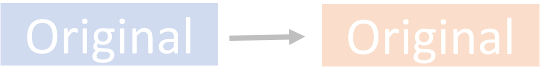
Original

[Nishimura et al., 2018]

Baseline | Multi-encoder NMT with back-translation

One-to-one NMT

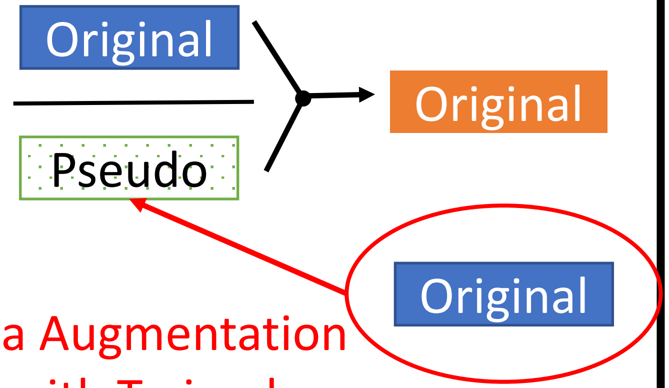
Standard NMT model from one source language to another target language



[Luong et al., 2015]

Multi-encoder NMT with back-translation

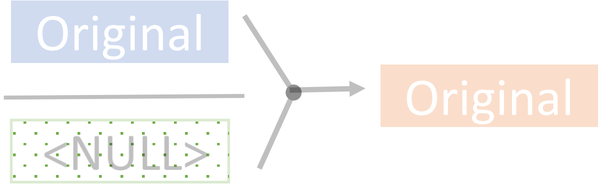
A multi encoder NMT system using pseudo-translation from English-to-X NMT



Data Augmentation with Trained one-to-one NMT English-to-X

Multi-encoder NMT with <NULL>

A multi-encoder NMT system using a special symbol <NULL>

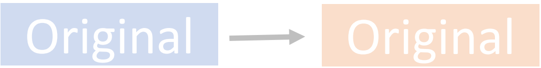


[Nishimura et al., 2018]

Baseline | Multi-encoder NMT with <NULL>

One-to-one NMT

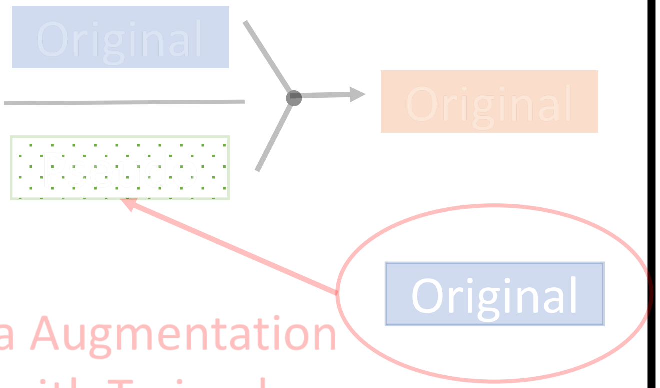
Standard NMT model from one source language to another target language



[Luong et al., 2015]

Multi-encoder NMT with back-translation

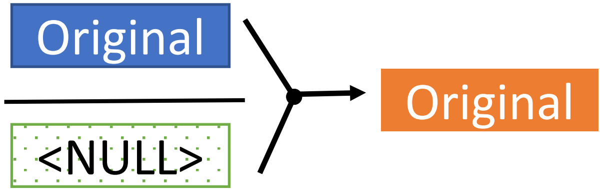
A multi encoder NMT system using pseudo-translation from English-to-X NMT



Data Augmentation with Trained one-to-one NMT English-to-X

Multi-encoder NMT with <NULL>

A multi-encoder NMT system using a special symbol <NULL>



[Nishimura et al., 2018]

Result

Result in BLEU

Pair	Trg	baseline method			proposed method		
		one-to-one (En-to-Trg)	multi-encoder NMT (fill up with symbol)	multi-encoder NMT (back-translation)	fill-in	fill-in and replace	fill-in and add
en-hr/sr	hr	20.21	28.18	27.57	29.17	29.37	**29.40**
	sr	16.42	23.85	22.73	24.41	**24.96**	24.15
en-sk/cs	sk	13.79	20.27	19.83	20.26	20.43	**20.59**
	cs	14.72	19.88	19.54	20.78	**20.90**	20.61
en-vi/id	vi	24.60	25.70	26.66	**26.73**	26.48	26.32
	id	24.89	**26.89**	26.34	26.40	25.73	26.21

Result | baseline vs proposed

Result in BLEU

Pair	Trg	baseline method			proposed method		
		one-to-one (En-to-Trg)	multi-encoder NMT (fill up with symbol)	multi-encoder NMT (back-translation)	fill-in	fill-in and replace	fill-in and add
en-hr/sr	hr	20.21	28.18	27.57	29.17	29.37	**29.40**
	sr	16.42	23.85	22.73	24.41	**24.96**	24.15
en-sk/cs	sk	13.79	20.27	19.83	20.26	20.43	**20.59**
	cs	14.72	19.88	19.54	20.78	**20.90**	20.61
en-vi/id	vi	24.60	25.70	26.66	**26.73**	26.48	26.32
	id	24.89	**26.89**	26.34	26.40	25.73	26.21

- en-hr/sr, en-sk/cs
  - Proposed methods > Baseline Method
  - proposed method is an effective way to use incomplete multilingual corpora

Result | baseline vs proposed

Result in BLEU

Pair	Trg	baseline method			proposed method		
		one-to-one (En-to-Trg)	multi-encoder NMT (fill up with symbol)	multi-encoder NMT (back-translation)	fill-in	fill-in and replace	fill-in and add
en-hr/sr	hr	20.21	28.18	27.57	29.17	29.37	**29.40**
	sr	16.42	23.85	22.73	24.41	**24.96**	24.15
en-sk/cs	sk	13.79	20.27	19.83	20.26	20.43	**20.59**
	cs	14.72	19.88	19.54	20.78	**20.90**	20.61
en-vi/id	vi	24.60	25.70	26.66	**26.73**	26.48	26.32
	id	24.89	**26.89**	26.34	26.40	25.73	26.21

- en-vi/id
  - Baseline Method > Proposed Method
  - The improvement by the use of multi-encoder NMT against one-to-one NMT in the baseline was small

Result | Three types of augmentation

Result in BLEU

Pair	Trg	baseline method			proposed method		
		one-to-one (En-to-Trg)	multi-encoder NMT (fill up with symbol)	multi-encoder NMT (back-translation)	fill-in	fill-in and replace	fill-in and add
en-hr/sr	hr	20.21	28.18	27.57	29.17	29.37	**29.40**
	sr	16.42	23.85	22.73	24.41	**24.96**	24.15
en-sk/cs	sk	13.79	20.27	19.83	20.26	20.43	**20.59**
	cs	14.72	19.88	19.54	20.78	**20.90**	20.61
en-vi/id	vi	24.60	25.70	26.66	**26.73**	26.48	26.32
	id	24.89	**26.89**	26.34	26.40	25.73	26.21

- There were almost no differences among three types of augmentation

Detail analysis →

We created three types of augmentation with one-to-one NMT output

Analysis | Three types of augmentation

Our expectation

The aggressive use (“fill-in and replace” and ”fill-in and add”) of low quality pseudo-translations

→ Contaminate the training data and to decrease the translation accuracy

We created three types of augmentation with one-to-one NMT output

Analysis | Three types of augmentation

Result in BLEU (Augment with one-to-one NMT)

Pair	Trg	Multi-encoder NMT (back-translation)		
		fill-in	fill-in and replace	fill-in and add
en-hr/sr	hr	**27.57**	24.05	24.79
	sr	**22.73**	17.77	22.02
en-sk/cs	sk	**19.83**	16.75	18.16
	cs	**19.54**	17.04	18.40
en-vi/id	vi	**26.66**	26.39	26.65
	id	26.34	23.90	**26.67**

→ large difference

→ large difference

→ few difference

- en-vi/id : there are few differences in three types of augmentation
  - one-to-one NMT was better than other language pairs

Analysis | Three types of augmentation

Result in BLEU
(Augment with one-to-one NMT)

Pair	Trg	Multi-encoder NMT (back-translation)		
		fill-in	fill-in and replace	fill-in and add
en-hr/sr	hr	**27.57**	24.05	24.79
	sr	**22.73**	17.77	22.02
en-sk/cs	sk	**19.83**	16.75	18.16
	cs	**19.54**	17.04	18.40
en-vi/id	vi	**26.66**	26.39	26.65
	id	26.34	23.90	**26.67**

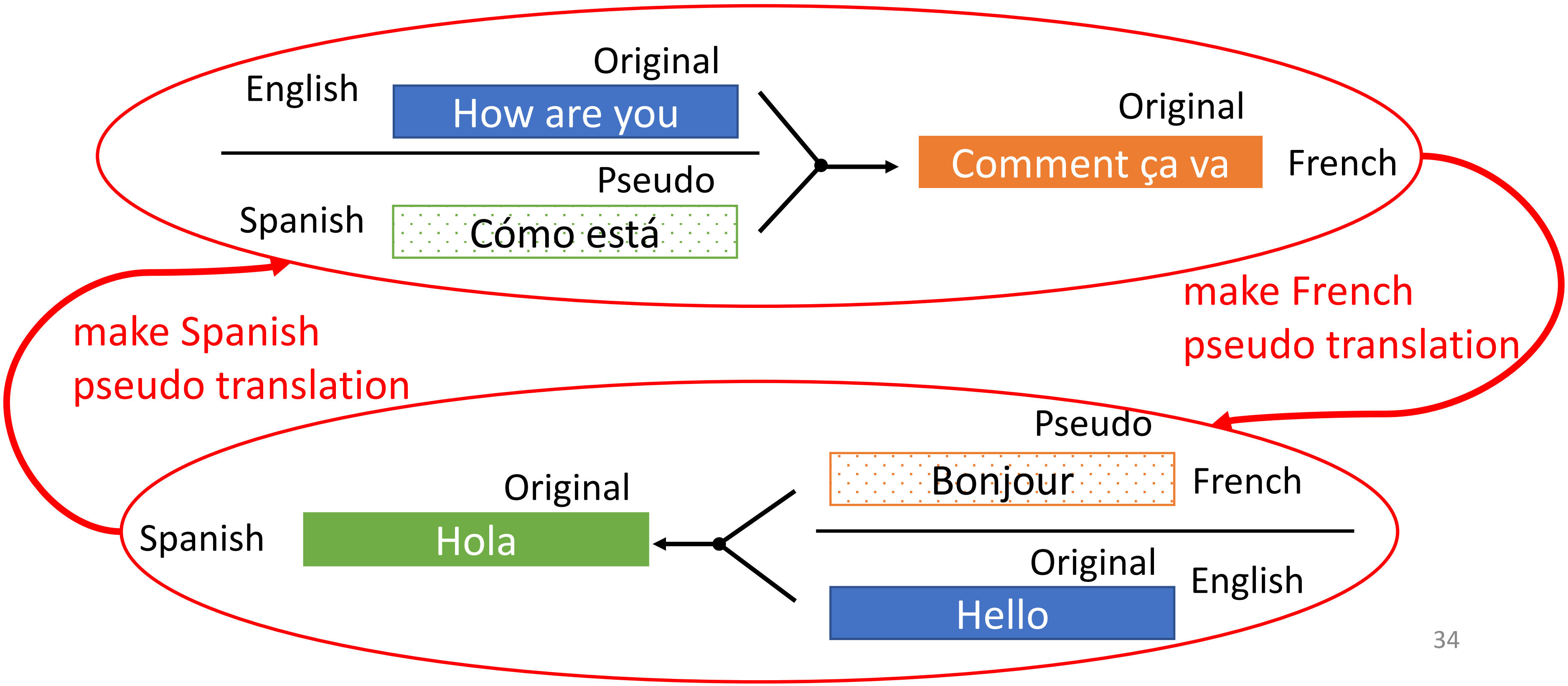
Train Data statistics

Pair	Trg	missing
en-hr/sr	hr	35564 (29.9%)
	sr	50203 (37.6%)
en-sk/cs	sk	58602 (57.7%)
	cs	17380 (29.0%)
en-vi/id	vi	87816 (54.5%)
	id	9424 (11.4%)

- Target=Indonesian : ”fill-in and add” got highest BLEU
  - much smaller fraction of the missing parts in Indonesian corpus

Analysis | Iterative Augmentation

- Update the multi-source NMT systems into the two target languages iteratively



Analysis | Iterative Augmentation

Result in BLEU (and BLEU gains compared to step1)

Pair	Trg	step1	step2	step3	step4
en-hr/sr	hr	29.17 (+0.00)	29.03 (-0.14)	29.10 (-0.07)	29.95 (-0.12)
	sr	24.41 (+0.00)	24.18 (-0.23)	24.17 (-0.24)	23.95 (-0.46)

- BLEU decreased gradually in every step
- We observed very similar results in the other language pairs

The iterative training may be introducing more noise

Analysis | Non-Parallelism

Example of the Serbian pseudo-translation

Type	Sentence
Original (En)	So **let me** conclude with just a remark to bring it back to the theme of choices.
Original (Sr)	Da zaključim jednom konstatacijom kojom se vraćam na temu izbora.
Pseudo (Sr)	**Dozvolite mi** da zaključim samo jednom opaskom, da se vratim na temu izbora.

- The Serbian original translation does not have a phrase corresponding to “let me”
- The Serbian pseudo translation have a phrase corresponding to “let me”

“fill-in and replace” or “fill-in and add” can be used to compensate for the missing information

Conclusion and future work

Conclusion

- Our research is the first study on how to handle incomplete corpora in multi-source NMT
- We proposed three types of augmentation
- Our proposed methods proved better than baseline systems, though results depend on the language pair

Future Work

- A set of three languages is that missing parts in the test sets could not be filled in, we will conduct experiments using more languages