Writing in Two Languages: Neural Machine Translation as an Assistive Bilingual Writing Tool

Jitao Xu

NetEase Youdao

Self-Introduction

Jitao Xu

I will join NetEase Youdao and Tsinghua University as a postdoctoral researcher.

PhD:

- Paris-Saclay University
- CNRS
- LISN (ex-LIMSI) & SYSTRAN
- Advisor: François Yvon

Writing in Two Languages: Neural Machine Translation as an Assistive Bilingual Writing Tool

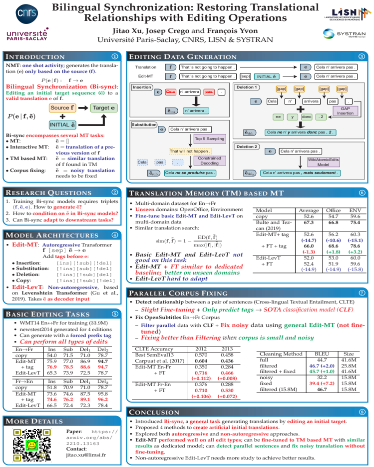
Jitao Xu

NetEase Youdao

Table of Contents

Using a foreign language in...

- Scientific activities
- International business
- Foreign videos



Request of booking confirmation letter for ACL2022

A2 ACL 2022 <acl2022@abbey.ie>
To: XU Jitao

Letter of support for visa app...
541 KB

Dear Jitao Xu,
Many thanks for your email and for sending us your details.

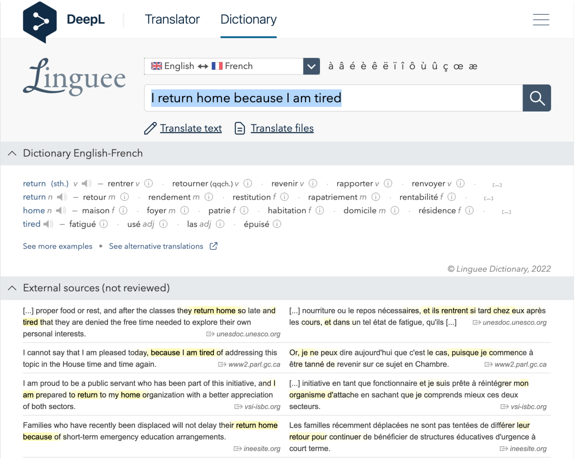
Please find attached your visa letter filled with all needed information.
Should you need any further assistance, please kindly let us know.
Kind regards,

Ms Solene Clement
Association for Computacional Linguistics
ACL 2022 Secretariat
E: acl2022@abbey.ie ¦ W: https://www.2022.aclweb.org/

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

Writing in a Foreign Language (L2)

- **NOT easy!**
- Fully relying on **NMT systems** is **not yet realistic**
  - May contain errors
  - Difficult to control
- Find help from **external resources** (dictionaries, terminologies, bilingual concordancers, etc.)
  - Interrupt the writing process



L2 Writing Assistance

System of Chen et al. (2012)

Type to translate

I *rentre à la maison* because I am tired.

English

I return home because I am tired.

- Bilingual composition
  - **Does not interrupt writing**
- L2 segments help to translate L1 segments (in native language)
  - **Better than direct translation**

L2 Writing Assistance

System of Chen et al. (2012)

Type to translate

I *rentre à la maison* because I am tired.

English

I return home because I am tired.

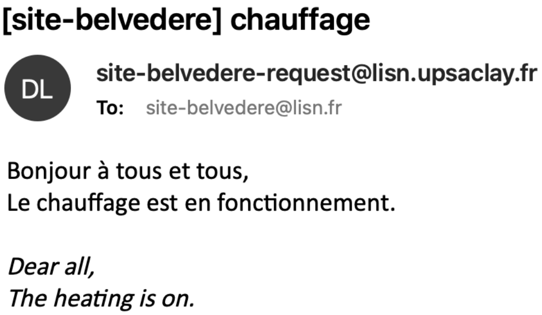
- Bilingual composition
  - **Does not interrupt writing**
- L2 segments help to translate L1 segments (in native language)
  - **Better than direct translation**
- **Only show full text in L2**
  - **Hard to evaluate**

Bilingual Writing

- **Bilingual composition**
- **Full texts in both L1 and L2**
  - Help **verify L2** with **corresponding L1** texts

Bilingual Writing

- **Bilingual composition**
- **Full texts in both L1 and L2**
    - Help **verify L2** with **corresponding L1** texts
    - Compose one sentence, obtain **synchronized bitext**



[site-belvedere] chauffage

DL

site-belvedere-request@lisn.upsaclay.fr

To: site-belvedere@lisn.fr

Bonjour à tous et tous,
Le chauffage est en fonctionnement.

Dear all,
The heating is on.

Bilingual Writing

Type to translate

I *rentre à la maison* because I am tired.

English

I return home because I am tired.

- Bilingual composition ✓
- Full texts in L1 and L2 ✗

Type to translate **French**

Je rentre à la maison parce que je suis fatigué.

English

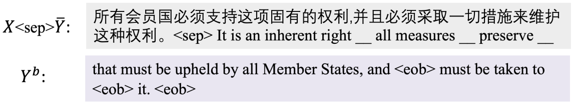
I return home because I am tired.

- Bilingual composition ✗
- Full texts in L1 and L2 ✓

Related Work

CAT system. Knowles and Koehn (2016)

Auto-completion. Li et al. (2021)



Bilingual text infilling. Xiao et al. (2022)

Source Sentence	他们也许并不知道这是一个 “假理财” 骗局，但也察觉到了诸多可疑之处，然而最终还是按照张颖的指使进行了违法违规操作。
Translation	They may not know this is a "fake financial management" scam, but also aware of many suspicious, and ultimately conduct illegal operations according to Zhang Ying's instructions.
Suggestions	1. suspects (s) 2. doubtful points (d p) 3. questionable points (q p)

Translation suggestion. Yang et al. (2022)

Our proposal No.1: Dual Decoding

Type to translate

I *rentre à la maison* because I am tired.

English

I return home because I am tired.

French

Je rentre à la maison parce que je suis fatigué.

- Mixed-language (MXL) composition — Bilingual composition ✓
- Display L1 and L2 in two boxes — Full texts in L1 and L2 ✓

Bilingual Writing Systems

Our proposal No.2: Bilingual Synchronization

- One language per box
- Both boxes allow composing — Bilingual composition ✓
- Display synchronized L1 and L2 — Full texts in L1 and L2 ✓

Bilingual Writing Systems

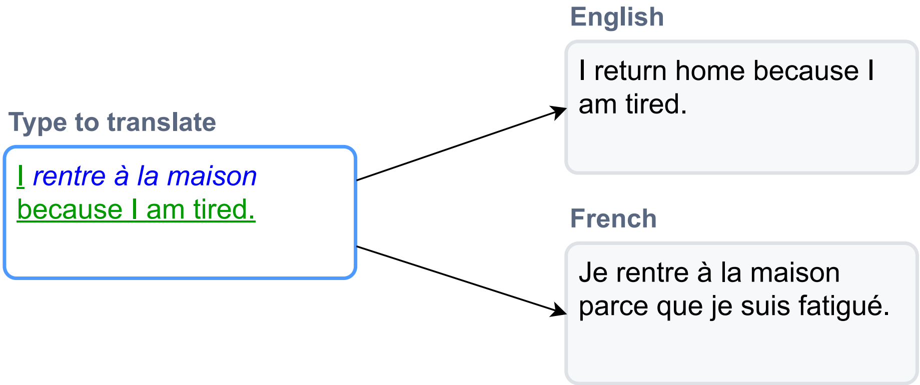
- Focused on **developing new techniques** for both proposed approaches
- Evaluated in **simulated** interactive **situations**

Research Questions:

- How to deal with MXL data? Do we need to **annotate words from different languages**?
- Is it possible to **simultaneously** generate **two targets** in one model?
- How to **efficiently** synchronize bitext?

Table of Contents

Dual Decoding



- Taking MXL sentence as input
- **Simultaneously** generating **consistent** translations in L1 and L2

Missing MXL Data

- Require triplets $(\mathbf{f}, \mathbf{e}^1, \mathbf{e}^2)$ for dual decoding
  - $\mathbf{f} \;\; = $ MXL sentence
  - $\mathbf{e}^1 = $ L1 sentence
  - $\mathbf{e}^2 = $ L2 sentence
- **Only have** parallel data e^1 **and** e^2

Missing MXL Data

- Require triplets $(\mathbf{f}, \mathbf{e}^1, \mathbf{e}^2)$ for dual decoding
  - $\mathbf{f}$ = MXL sentence
  - $\mathbf{e}^1$ = L1 sentence
  - $\mathbf{e}^2$ = L2 sentence
- **Only have** parallel data $\mathbf{e}^1$ **and** $\mathbf{e}^2$
- **Generate synthetic MXL data** $\mathbf{f}$ from $\mathbf{e}^1$ and $\mathbf{e}^2$
  - Main language: preserving the **sentence structure**
  - Secondary language: **inserted segments**
  - **Replace main** segments with **secondary** ones

MXL Data Generation

Alignment units

In Oregon , planners are experimenting with giving drivers different choices .
Dans l'Orégon , les planificateurs tentent l'expérience en offrant aux automobilistes différents choix .

- Select **the main language** and **number of replacements** r according to:

$$P(r = k) = \frac{1}{2^{k+1}} \quad \forall k = 1, \ldots, R$$

- Make sure r smaller than half of either side's length

$$r = \min(\frac{|S|}{2}, \frac{|T|}{2}, r)$$

- **Randomly replace** r main units with secondary ones

MXL Data Generation

Generated MXL sentences

Main	In Oregon , planners are experimenting with giving drivers different choices .
$r = 1$	**Dans** Oregon , planners are experimenting with giving drivers different choices .
$r = 2$	**Dans** Oregon , **les planificateurs** are experimenting with giving drivers different choices .
$r = 3$	**Dans** Oregon , **les planificateurs** are experimenting **en offrant aux** drivers different choices .
Secondary	Dans l'Orégon , les planificateurs tentent l'expérience en offrant aux automobilistes différents choix .

Model Architecture

- MXL data ✓
- How to **simultaneously generate consistent L1 and L2?**

Model Architecture

- MXL data ✓
- How to **simultaneously generate consistent L1 and L2?**

Dual Decoder Model

Dual Decoder Model

Simultaneously translating a source $\mathbf{f}$ into two targets $\mathbf{e}^1$ and $\mathbf{e}^2$:

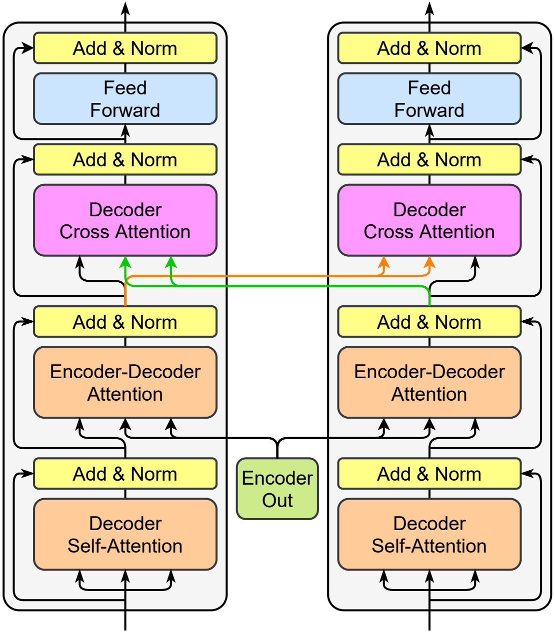
$$P(\mathbf{e}^1, \mathbf{e}^2|\mathbf{f}) = \prod_{t=1}^{T} P(\mathbf{e}_t^1, \mathbf{e}_t^2|\mathbf{f}, \mathbf{e}_{<t}^1, \mathbf{e}_{<t}^2)$$

`dual` $$P(\mathbf{e}^1, \mathbf{e}^2|\mathbf{f}) = \prod_{t=1}^{T} P(\mathbf{e}_t^1|\mathbf{f}, \mathbf{e}_{<t}^1, \mathbf{e}_{<t}^2) \times P(\mathbf{e}_t^2|\mathbf{f}, \mathbf{e}_{<t}^1, \mathbf{e}_{<t}^2)$$

$$P(\mathbf{e}^1, \mathbf{e}^2|\mathbf{f}) = \prod_{t=1}^{T} P(\mathbf{e}_t^1|\mathbf{f}, \mathbf{e}_{<t}^1) P(\mathbf{e}_t^2|\mathbf{f}, \mathbf{e}_{<t}^2)$$

- One shared encoder, two **synchronized decoders**
- Synchronous decoding ($\mathbf{e}_t^1$ and $\mathbf{e}_t^2$) is performed **simultaneously** at each step

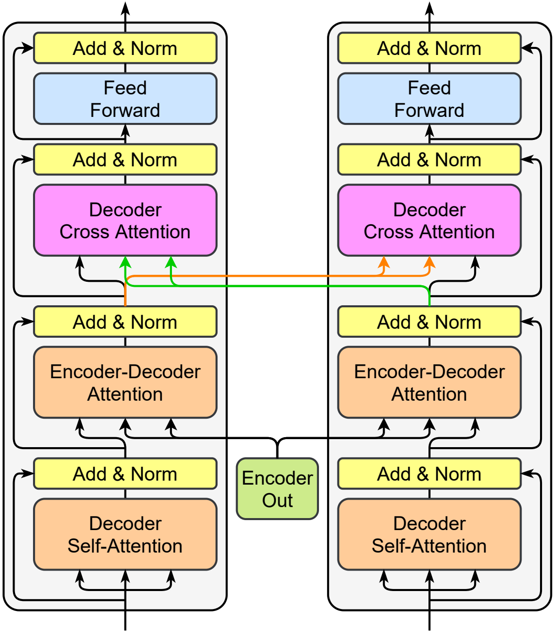
Dual Decoder Model



Hidden states of layer l as H_l^1 and H_l^2:

$$H_{l+1}^1 = \mathrm{Attention}(H_l^1, H_l^2, H_l^2)$$
$$H_{l+1}^2 = \mathrm{Attention}(H_l^2, H_l^1, H_l^1)$$

Dual Decoder Model



Hidden states of layer l as H_l^1 and H_l^2:

$$H_{l+1}^1 = \text{Attention}(H_l^1, H_l^2, H_l^2)$$
$$H_{l+1}^2 = \text{Attention}(H_l^2, H_l^1, H_l^1)$$

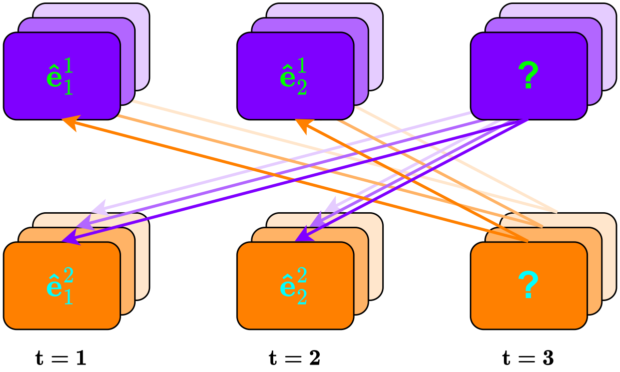
Training and with **a combined loss**:

$$L(\theta) = \sum_D \Big(\sum_{t=1}^{|\mathbf{e}^1|} \log P(\mathbf{e}_t^1 | \mathbf{e}_{<t}^1, \mathbf{e}_{<t}^2, \mathbf{f}, \theta) + \sum_{t=1}^{|\mathbf{e}^2|} \log P(\mathbf{e}_t^2 | \mathbf{e}_{<t}^2, \mathbf{e}_{<t}^1, \mathbf{f}, \theta) \Big)$$

Decoding with Decoder Cross Attention

Dual beam search:

- Each candidate only **attends to one candidate** from the other decoder



Experimental Settings

- Data:
  Training: WMT14 En-Fr & WMT13 En-Es
  Test: **`newstest2014`** for En-Fr, **`newstest2013`** for En-Es
  Generate **synthetic** MXL `newstest2014` and `newstest2013`
- Models:
  - `dual`: Our dual decoder model
  - 3 **MXL** baselines:
    `base`: **Two separate** Transformers *e.g.* MXL-En + MXL-Fr
    `multi`: **One multilingual model** for *e.g.* MXL-En & MXL-Fr
    `indep`: One encoder, **two independent decoders** with a joint loss
  - 2 **monolingual** baselines:
    `base-mono`: *e.g.* En-Fr + Fr-En
    `bilingual`: *e.g.* En-Fr & Fr-En

Results

- dual comparable to bilingual on **monolingual** sentence
- dual similar to base on **MXL**

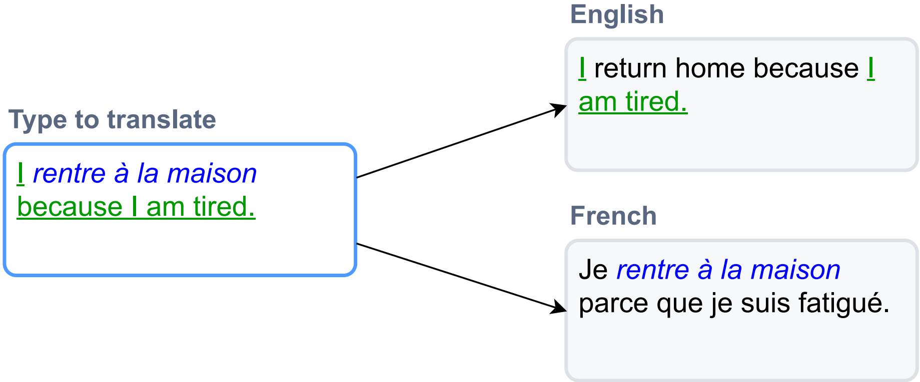
BLEU	newstest2014		mxl-newstest2014	
Direction	En-Fr	Fr-En	MXL-Fr	MXL-En
copy	-	-	50.0	46.5
base-mono	37.6	35.2	45.0	61.3
bilingual	**36.1**	**34.0**	46.3	59.4
base	36.5	34.1	**67.4**	**67.8**
multi	34.6	32.3	66.4	65.7
indep	35.9	34.0	67.3	67.7
dual	**36.0**	**33.9**	**67.5**	**67.7**

Results

- dual comparable to bilingual on **monolingual** sentence
- dual similar to base on **MXL**
- dual better than multi

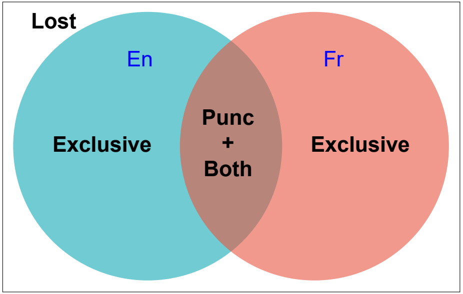
BLEU	newstest2014		mxl-newstest2014	
Direction	En-Fr	Fr-En	MXL-Fr	MXL-En
copy	-	-	50.0	46.5
base-mono	37.6	35.2	45.0	61.3
bilingual	36.1	34.0	46.3	59.4
base	36.5	34.1	67.4	67.8
multi	34.6	32.3	**66.4**	**65.7**
indep	35.9	34.0	67.3	67.7
dual	36.0	33.9	**67.5**	**67.7**

Copy Constraint



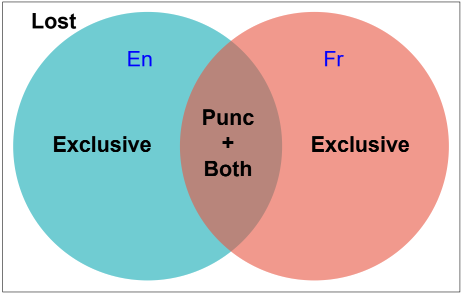
- **User-composed** texts should be **preserved** in the two translations
- **All words** in MXL should appear in **at least one output**

Copy Constraint



Copy Constraint

	En-Fr			
Model	Exclusive	Punc	Both	Lost
reference	81.56	10.34	8.10	0.00
base	**79.14**	11.29	**8.85**	0.72
multi	78.66	11.27	9.22	0.85
indep	78.86	11.35	9.13	0.67
dual	**78.90**	11.32	**9.17**	**0.61**



- **Distinguish one language from another**
- Different translation choices
- `dual` has **fewer lost tokens**

L2 Writing Assistant Task (SemEval 2014)

Example (L1=French, L2=English)

Input: *"I **rentre à la maison** because I am tired."*
Reference: *"I **return home** because I am tired."*

- **Translating L1 fragments** in **L2 contexts**
- A more realistic task
- Direct zero-shot inference on this task

L2 Writing Assistant Task

Fr-En	Accuracy	Word Accuracy	Recall
UEdin-run1	0.733	0.824	1.0
UEdin-run2	0.731	0.821	1.0
UEdin-run3	0.723	0.816	1.0
CNRC-run1	0.556	0.694	1.0
`dual`	0.602	0.723	0.998

En-Es	Accuracy	Word Accuracy	Recall
UEdin-run2	0.755	0.827	1.0
UEdin-run1	0.753	0.827	1.0
UEdin-run3	0.745	0.820	1.0
`dual`	**0.787**	**0.854**	1.0

- **Zero-shot inference**
- 4th place for Fr-En
- **State-of-the-art** for En-Es

More Applications with Dual Decoder Model

Source	I could do that again if you want .	
L2R	Je peux le refaire si vous le voulez .	**Bidirectional decoding**
R2L	. voulez le vous si refaire le peux Je	
polite	Ich kann das noch mal machen , wenn Sie wollen .	**Multi-style Decoding**
informal	Ich kann das noch mal machen , wenn du willst .	

Transcript	**i 'm** combining specific types of signals **the** mimic how our body **response** to **in an injury** to help us regenerate	
Caption	**I'm** combining specific types of signals **[eob]** **that** mimic how our body **responds** to **injury** **[eol]** to help us regenerate**.** **[eob]**	**Multilingual**
Subtitle	Je combine différents types de signaux **[eob]** qui imitent la réponse du corps **[eol]** aux blessures pour nous aider à guérir. **[eob]**	**subtitling**

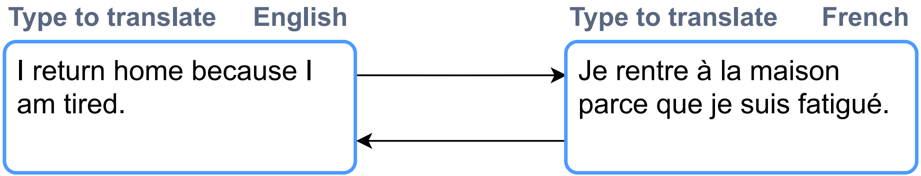
- Applied to other tasks. **Mitigated exposure bias** problem. Obtained **similar or better** performance with **higher consistency** between outputs.

Summary of Dual Decoding

- **Simultaneously** translate MXL into L1 and L2
- Generate **synthetic MXL** data
- Proposed **dual decoder** model, **simultaneously** generating pairs of **consistent** translations
- Very **few lost tokens**
- Implicit **language identification** ability
- Zero-shot inference on **realistic** L2 writing assistant task

Table of Contents

Bilingual Synchronization



- Allow composing on both sides
- Keep texts in L1 and L2 **synchronized**
- Make **small changes** through revision

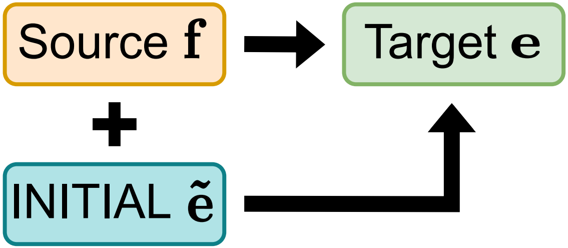
Bilingual Synchronization (Bi-sync)

Given:

- $\mathbf{f}$: a source
- $\tilde{\mathbf{e}}$: an initial target, **small differences** to $\mathbf{e}$

Find $\mathbf{e}$: **translation of** $\mathbf{f}$, by **editing** $\tilde{\mathbf{e}}$

$$P(\mathbf{e} \mid \mathbf{f}, \tilde{\mathbf{e}})$$



A General Task

Bi-sync **encompasses several MT tasks**:

Bilingual writing:	$\tilde{\mathrm{e}}$ = **translation of a previous version** of f
Translation Memory (TM) based MT:	$\tilde{\mathrm{e}}$ = **similar translation** of f found in TM
Parallel corpus fixing:	$\tilde{\mathrm{e}}$ = **noisy translation** needs to be fixed
Automatic post-editing:	$\tilde{\mathrm{e}}$ = **MT output** to edit
MT:	$\tilde{\mathrm{e}} = []$

A General Task

Bi-sync **encompasses several MT tasks**:

Bilingual writing:	$\tilde{e}$ = **translation of a previous version** of f
Translation Memory (TM) based MT:	$\tilde{e}$ = **similar translation** of f found in TM
Parallel corpus fixing:	$\tilde{e}$ = **noisy translation** needs to be fixed
Automatic post-editing:	$\tilde{e}$ = **MT output** to edit
MT:	$\tilde{e} = []$

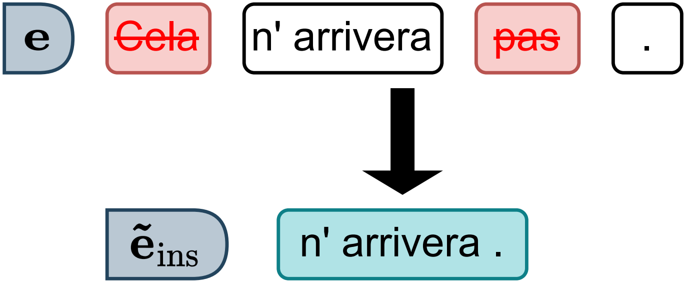
Generating Training Editing Data

- Require triplets $(\mathbf{f}, \tilde{\mathbf{e}}, \mathbf{e})$
- **Small edits** between $\tilde{\mathrm{e}}$ and e
- **Only have** parallel data f **and** e

Generating Training Editing Data

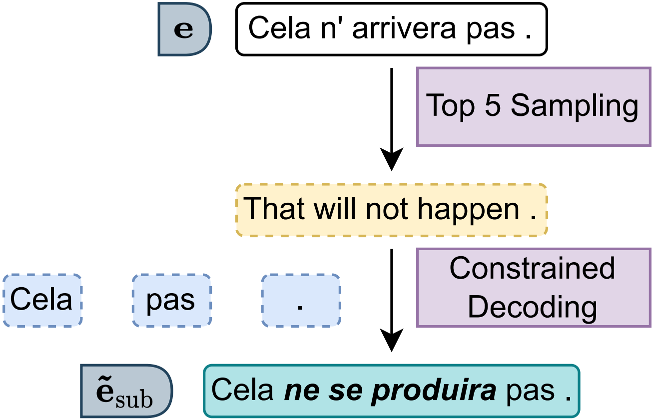
- Require triplets $(\mathbf{f}, \tilde{\mathbf{e}}, \mathbf{e})$
- **Small edits** between $\tilde{\mathrm{e}}$ and e
- **Only have** parallel data f **and** e
- **Decompose editions** as basic types: Insertion, Substitution, Deletion
- Generate **synthetic** $\tilde{\mathrm{e}}$ for **each editing type**

Insertion



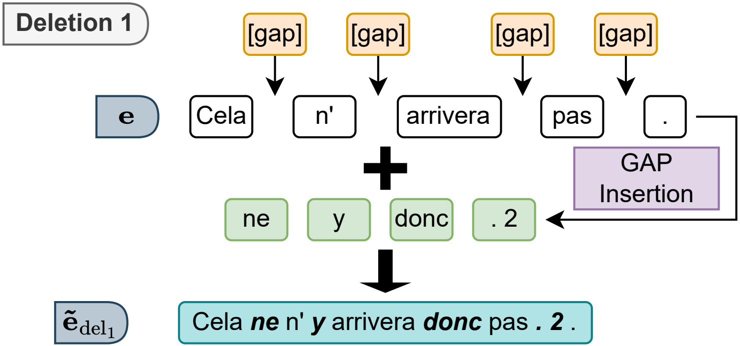
- **Randomly drop tokens** from e
- Keep at least half of e

Substitution



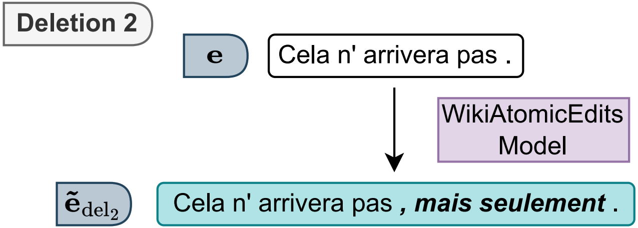
- Round trip translation with constrained decoding
- Back translate $\mathbf{e} \rightarrow \mathbf{f}^*$ with **top-5 sampling**
- $\mathbf{f}^* \rightarrow \tilde{\mathbf{e}}_{\text{sub}}$ with **lexical constrained decoding**
- Half of $\mathbf{e}$ as constraints, substitute the other half

Deletion



Randomly combine

$\tilde{\mathrm{e}}_{\mathrm{del}_1}$ and $\tilde{\mathrm{e}}_{\mathrm{del}_2}$ as $\tilde{\mathrm{e}}_{\mathrm{del}}$



Copy and Translate

Copy

- **Detect parallelism** between $\mathbf{f}$ and $\tilde{\mathbf{e}}$
- **Do not change** anything if already parallel
- $\tilde{\mathbf{e}}_{\text{cp}} = \mathbf{e}$

Editing and translation

- Final $\tilde{\mathbf{e}}$: **random combination** of $\tilde{\mathbf{e}}_{\text{ins}}$, $\tilde{\mathbf{e}}_{\text{sub}}$, $\tilde{\mathbf{e}}_{\text{del}}$ and $\tilde{\mathbf{e}}_{\text{cp}}$
- Combine editing data $(\mathbf{f}, \tilde{\mathbf{e}}, \mathbf{e})$ and translation data $(\mathbf{f}, \mathbf{e})$
- Keep translation ability

Model Architecture

- Editing data $\tilde{e}$ ✓
- How to **condition on $\tilde{e}$?**

Model Architecture

- Editing data $\tilde{e}$ ✓
- How to **condition on $\tilde{e}$?**

Two approaches:
Autoregressive and **non-autoregressive**

- Non-autoregressive model is **more efficient**

Edit-MT

Translation **f** That 's not going to happen . → **e** Cela n' arrivera pas .

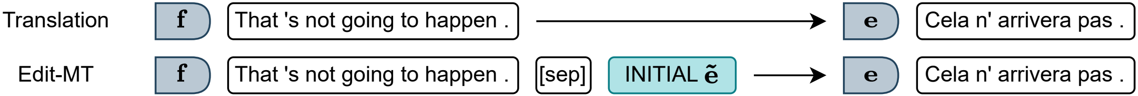
Edit-MT **f** That 's not going to happen . [sep] INITIAL $\tilde{\mathbf{e}}$ → **e** Cela n' arrivera pas .

- **Autoregressive**, similar to Bulte and Tezcan (2019)
- **Prefix editing tags** on target side

Tagging scheme:

Insertion:	`[ins] [!sub] [!del]`
Substitution:	`[!ins] [sub] [!del]`
Deletion:	`[!ins] [!sub] [del]`
Copy:	`[!ins] [!sub] [!del]`

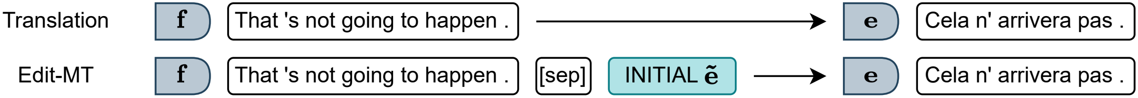
Edit-MT



Inference with tag:

- Direct inference: predict **tag** + $\mathbf{e}$ (**Tags unknown**)
- Prefix decoding: **forced prefix tag** + **predict** $\mathbf{e}$ (**Tags known**)

Edit-MT

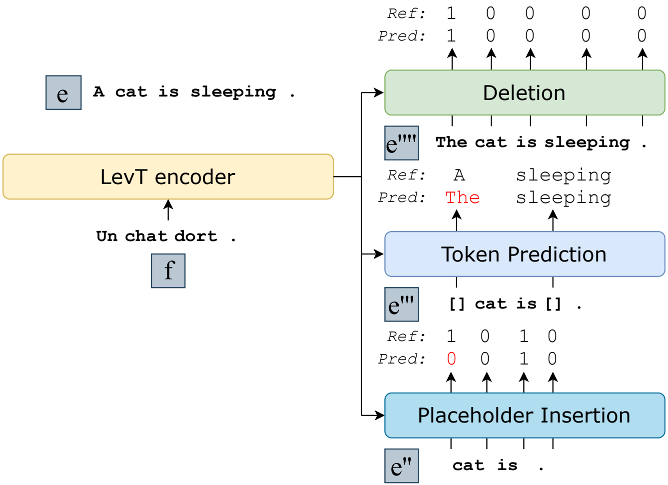


Inference with tag:

- Direct inference: predict **tag** + e (**Tags unknown**)
- Prefix decoding: **forced prefix tag** + **predict** e (**Tags known**)

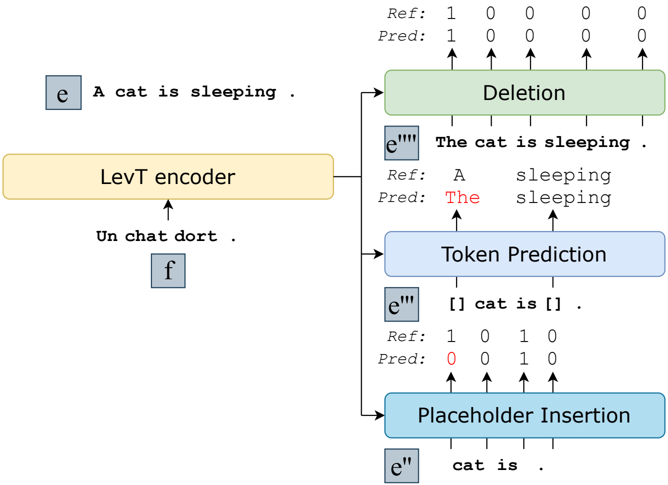
Autoregressive model **does not really make edits to** $\tilde{\mathrm{e}}$

Levenshtein Transformer (LevT)

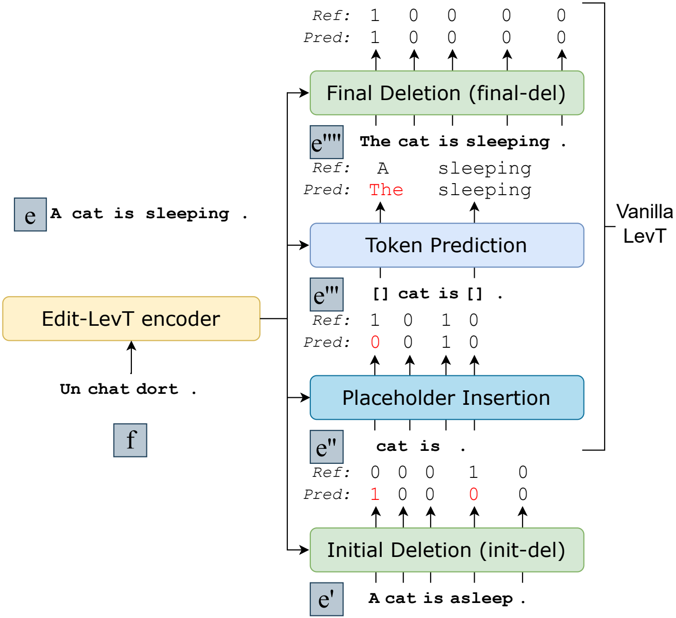


- **Non-autoregressive**
- Perform **edits** to a sentence
- Iterative refinement decoding

Levenshtein Transformer (LevT)



- **Non-autoregressive**
- Perform **edits** to a sentence
- Iterative refinement decoding
  - Always starts from **empty**
- **Only delete prediction errors**



- **Non-autoregressive**, based on LevT
  - LevT only deletes **prediction errors**
- Need to **remove unrelated tokens** from $\tilde{\mathrm{e}}$
- Add an **initial deletion**
- $\mathrm{e}' = \tilde{\mathrm{e}}$
- Does not change inference

Experimental Settings

- Data:
  Training: WMT14 En-Fr
  Test: **`newstest2014`**
  Generate **synthetic** $\tilde{\mathrm{e}}$ for `newstest2014`
- Models:
  - **Edit-MT**: Our autoregressive model
  - **Edit-LevT**: Our non-autoregressive model
  - 2 baseline settings:
    `copy`: use $\tilde{\mathrm{e}}$ as output
    vanilla LevT: No initial deletion

Results for Basic Edits

Baseline translation 36.4 BLEU Avg. Edit-MT-1.2 BLEU. Edit-LevT-7.7 BLEU

En-Fr	Ins	Sub	Del_1	Del_2
copy	54.0	71.5	71.0	78.7
Edit-MT	**75.9**	**77.0**	**86.9**	**94.7**
+ tag	76.9	78.5	88.6	94.7
LevT	65.3	73.9	72.5	78.7
Edit-LevT	**72.6**	**76.3**	**81.9**	**92.2**

Fr-En	Ins	Sub	Del_1	Del_2
copy	51.8	70.9	71.0	78.7
Edit-MT	**73.6**	**74.6**	**87.5**	**95.8**
+ tag	74.6	76.2	89.1	96.2
LevT	66.5	72.4	72.3	78.4
Edit-LevT	**70.7**	**74.1**	**82.8**	**92.7**

- Edit-MT and Edit-LevT performs **all types of edit**

Results for Basic Edits

Baseline translation 36.4 BLEU Avg. Edit-MT-1.2 BLEU. Edit-LevT-7.7 BLEU

En-Fr	Ins	Sub	Del_1	Del_2
copy	54.0	71.5	71.0	78.7
Edit-MT	75.9	77.0	86.9	94.7
+ tag	**76.9**	**78.5**	**88.6**	**94.7**
LevT	65.3	73.9	72.5	78.7
Edit-LevT	72.6	76.3	81.9	92.2

Fr-En	Ins	Sub	Del_1	Del_2
copy	51.8	70.9	71.0	78.7
Edit-MT	73.6	74.6	87.5	95.8
+ tag	**74.6**	**76.2**	**89.1**	**96.2**
LevT	66.5	72.4	72.3	78.4
Edit-LevT	70.7	74.1	82.8	92.7

- Edit-MT and Edit-LevT performs **all types of edit**
- Edit-MT **+ tag works best**

Results for Basic Edits

Baseline translation 36.4 BLEU Avg. Edit-MT-1.2 BLEU. **Edit-LevT-7.7 BLEU**

En-Fr	Ins	Sub	Del$_1$	Del$_2$
copy	54.0	71.5	71.0	78.7
Edit-MT	**75.9**	**77.0**	**86.9**	**94.7**
+ tag	76.9	78.5	88.6	94.7
LevT	65.3	73.9	72.5	78.7
Edit-LevT	**72.6**	**76.3**	**81.9**	**92.2**

Fr-En	Ins	Sub	Del$_1$	Del$_2$
copy	51.8	70.9	71.0	78.7
Edit-MT	**73.6**	**74.6**	**87.5**	**95.8**
+ tag	74.6	76.2	89.1	96.2
LevT	66.5	72.4	72.3	78.4
Edit-LevT	**70.7**	**74.1**	**82.8**	**92.7**

- Edit-MT and Edit-LevT performs **all types of edit**
- Edit-MT **+ tag works best**
- Edit-LevT **close to Edit-MT**, depends on operation type
- Edit-LevT $3\times$ **faster** than Edit-MT

Multilingual Results

En-Fr	Ins	Sub	Del_1	Del_2
copy	54.0	71.5	71.0	78.7
Edit-MT	75.9	77.0	86.9	94.7
+ tag	76.9	78.5	88.6	94.7
`multi` Edit-MT	**75.5**	**77.2**	**86.9**	**94.7**
+ tag	**76.2**	**78.1**	**88.5**	**94.9**
Edit-LevT	72.6	76.3	81.9	92.2
`multi` Edit-LevT	**72.4**	**76.3**	**83.0**	**92.4**

- **Combine data** in both directions
- **No performance loss** for **multilingual** models
- Do not distinguish a target language
- **real BILINGUAL synchronization**

More Applications with Bi-sync Models

Bi-sync **encompasses several MT tasks**:

Bilingual writing:	$\tilde{e} =$ **translation of a previous version** of f
Translation Memory (TM) based MT:	$\tilde{e} =$ **similar translation** of f found in TM
Parallel corpus fixing:	$\tilde{e} =$ **noisy translation** needs to be fixed
Automatic post-editing:	$\tilde{e} =$ **MT output** to edit
MT:	$\tilde{e} = []$

- **Fine-tuning** on downstream tasks
- **Similar or even better** performance than **dedicated systems**

More Applications with Bi-sync Models

Bi-sync **encompasses several MT tasks**:

Bilingual writing:	$\tilde{e}$ = **translation of a previous version** of f
Translation Memory (TM) based MT:	$\tilde{e}$ = **similar translation** of f found in TM
Parallel corpus fixing:	$\tilde{e}$ = **noisy translation** needs to be fixed
Automatic post-editing:	$\tilde{e}$ = **MT output** to edit
MT:	$\tilde{e} = []$

- **Fine-tuning** on downstream tasks
- **Similar or even better** performance than **dedicated systems**

TM-based MT

- Find a **similar translation** of $\mathbf{f}$ from TM
- Make use of similar translation
- **Multiple edit operations** in one sentence

Experimental Settings:

- Multi-domain (11) data for En-Fr
- **Unseen domains**: OpenOffice and ENV
- **Zero-shot inference** & **fine-tuning**

Results for TM-based MT

BLEU	All 11	Office	ENV
copy	52.6	54.7	59.6
Bulte and Tezcan (2019)	**67.3**	**66.8**	**75.4**
Edit-MT+ tag	52.6	56.2	60.3
+ FT + tag	**66.0**	**68.6**	**78.6**
Edit-LevT	51.4	54.4	59.8
+ FT	**61.5**	**62.2**	**75.1**

- **Zero-shot** inference **does not work**
- **Fine-tuning** works well
- **Edit-MT + FT** similar to Bulte and Tezcan (2019)
- **Edit-LevT benefits** from fine-tuning

Summary of Bilingual Synchronization

- Define **Bi-sync** task
- Generate editing data **for each type**
- Propose **autoregressive** and **non-autoregressive** models to perform Bi-sync
- Good performance for **each editing type**
- Experiment with **multilingual approach**
- **Applicable to downstream tasks** like TM-based MT

Table of Contents

Conclusion

- Targeting **bilingual writing**
- Two approaches: **Dual Decoding** and **Bilingual Synchronization**

Conclusion

- Targeting **bilingual writing**
- Two approaches: **Dual Decoding** and **Bilingual Synchronization**

Dual decoding:

- **Simultaneously** generate L1 and L2 from **MXL**
- Generated **synthetic MXL**
- Proposed dual decoder model

Bilingual synchronization:

- Obtain translation of source by **editing an initial target**
- Generated editing data
- Proposed autoregressive and non-autoregressive approach

Conclusion

- Targeting **bilingual writing**
- Two approaches: **Dual Decoding** and **Bilingual Synchronization**

Dual decoding:

- **Simultaneously** generate L1 and L2 from **MXL**
- Generated **synthetic MXL**
- Proposed dual decoder model

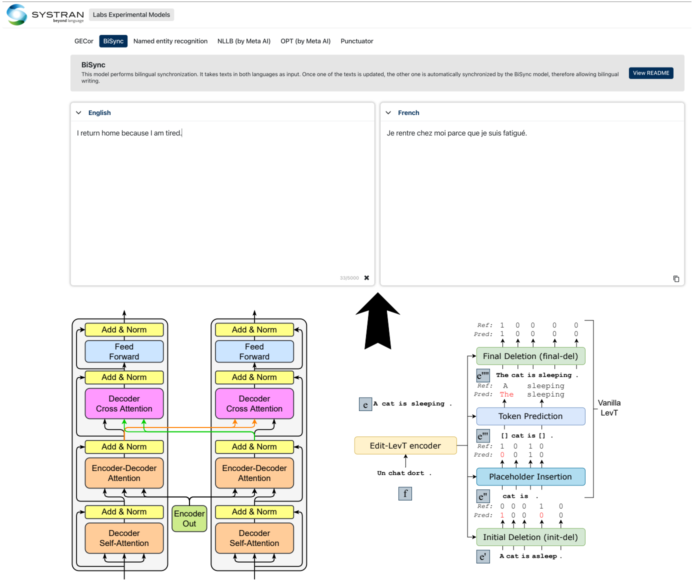
Bilingual synchronization:

- Obtain translation of source by **editing an initial target**
- Generated editing data
- Proposed autoregressive and non-autoregressive approach

- Both are **general framework**
- Applicable to other tasks with good performance

Future Perspectives

- Interface design and development
- Conduct user studies
- Evaluate the efficiency of bilingual writing tools in real scenarios
- Compare dual decoding with bilingual synchronization



Thank you!

References

Bram Bulte and Arda Tezcan. 2019. Neural fuzzy repair: Integrating fuzzy matches into neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1800–1809, Florence, Italy. Association for Computational Linguistics.

Mei-Hua Chen, Shih-Ting Huang, Hung-Ting Hsieh, Ting-Hui Kao, and Jason S. Chang. 2012. FLOW: A first-language-oriented writing assistant system. In *Proceedings of the ACL 2012 System Demonstrations*, pages 157–162, Jeju Island, Korea. Association for Computational Linguistics.

Rebecca Knowles and Philipp Koehn. 2016. Neural interactive translation prediction. In *Conferences of the Association for Machine Translation in the Americas: MT Researchers' Track*, pages 107–120, Austin, TX, USA. The Association for Machine Translation in the Americas.

Huayang Li, Lemao Liu, Guoping Huang, and Shuming Shi. 2021. GWLAN: General word-level AutocompletioN for computer-aided translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4792–4802, Online. Association for Computational Linguistics.

Yanling Xiao, Lemao Liu, Guoping Huang, Qu Cui, Shujian Huang, Shuming Shi, and Jiajun Chen. 2022. BiTIIMT: A bilingual text-infilling method for interactive machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1958–1969, Dublin, Ireland. Association for Computational Linguistics.

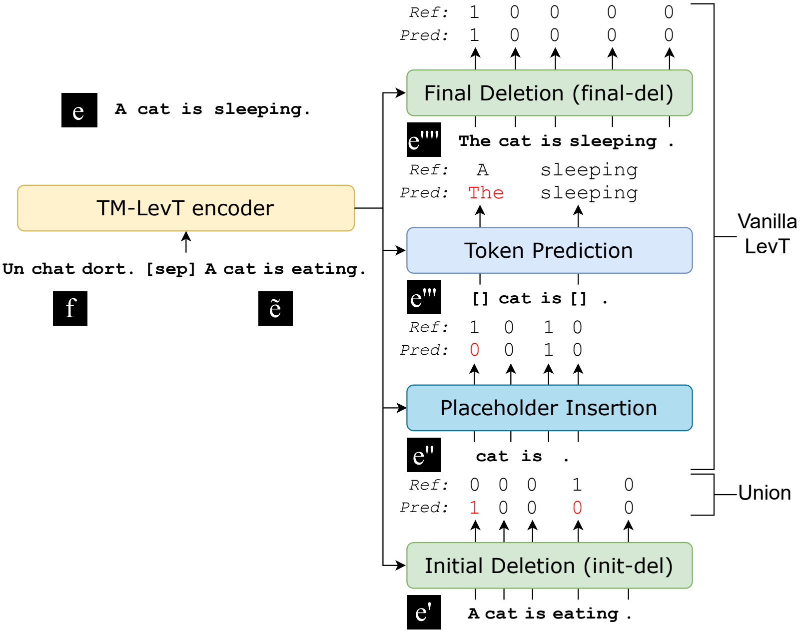
Zhen Yang, Fandong Meng, Yingxue Zhang, Ernan Li, and Jie Zhou. 2022. WeTS: A benchmark for translation suggestion. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5278–5290, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Analysis of Different Edit Types

BLEU	=	I	S	D	I+S	I+D	S+D	I+S+D	All
copy	100.0	72.0	67.9	75.4	32.5	69.8	34.0	47.3	52.6
Bulte and Tezcan (2019)	91.6	80.6	86.6	82.9	50.0	67.4	58.4	63.0	67.3
Edit-MT+ FT + tag	91.6	**79.7**	**84.6**	**85.8**	**48.3**	**69.9**	**57.6**	**60.8**	66.0
Edit-LevT + FT	**94.1**	77.5	81.1	81.4	41.8	67.7	52.0	56.7	61.5

- **Edit-MT + FT performs better on single edit type**
- **Edit-LevT + FT** good at **detecting parallelism**

Further Study of TM-based NAT



- **Extend f with similar translation ẽ**
- ẽ always accessible on source
- $e' = \tilde{e}$

Results

	sim > 0.6		sim $\in [0.4, 0.6]$	
BLEU	w/o TM	w/ TM	w/o TM	w/ TM
copy	-	52.6	-	34.5
Bulte and Tezcan (2019)	**51.2**	**67.1**	**46.1**	**55.7**
LevT	46.5	60.4	40.8	49.3
+tgt TM	-	52.8	-	35.0
Edit-LevT	**52.6**	**65.9**	**45.7**	**53.3**

- Edit-LevT **similar to autoregressive** baseline **with and without** TM
- Training with TMs **helps regular MT** for Edit-LevT

Knowledge Distillation

	$\text{sim} > 0.6$		$\text{sim} \in [0.4, 0.6]$	
BLEU	w/o TM	w/ TM	w/o TM	w/ TM
copy	-	52.6	-	34.5
Teacher	56.7	-	49.6	-
Edit-LevT	**52.6**	65.9	**45.7**	53.3
+KD	**54.3**	57.1	**47.6**	49.3
+KD TM	53.8	56.0	47.3	48.5

- KD helps regular translation

Knowledge Distillation

	sim > 0.6		sim $\in [0.4, 0.6]$	
BLEU	w/o TM	w/ TM	w/o TM	w/ TM
copy	-	52.6	-	34.5
Teacher	56.7	-	49.6	-
Edit-LevT	52.6	**65.9**	45.7	**53.3**
+KD	54.3	**57.1**	47.6	**49.3**
+KD TM	53.8	56.0	47.3	48.5

- KD helps regular translation
- KD **does not help** when **using TMs**

Knowledge Distillation

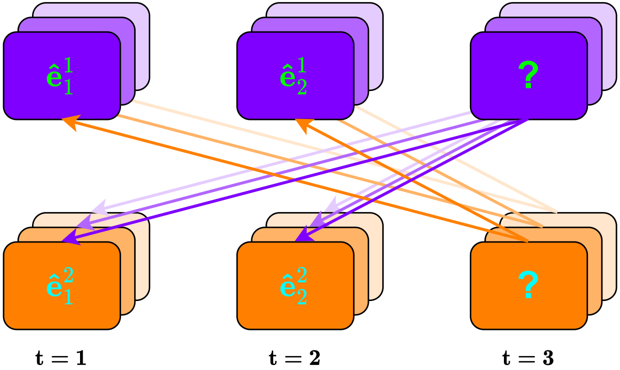
	sim > 0.6		sim $\in [0.4, 0.6]$	
BLEU	w/o TM	w/ TM	w/o TM	w/ TM
copy	-	52.6	-	34.5
Teacher	**56.7**	-	**49.6**	-
Edit-LevT	52.6	65.9	45.7	53.3
+KD	54.3	**57.1**	47.6	**49.3**
+KD TM	53.8	56.0	47.3	48.5

- KD helps regular translation
- KD **does not help** when **using TMs**
- Performance with KD **limited to teacher**

Decoding with Decoder Cross Attention

Dual beam search:

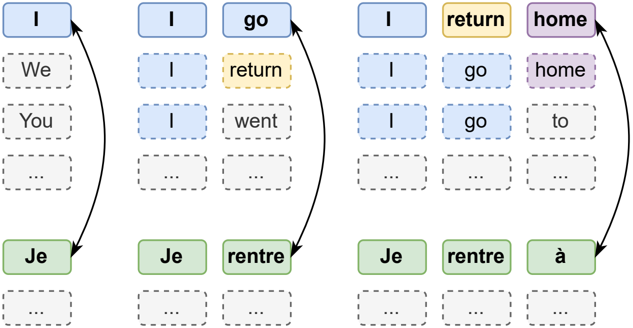
- Each candidate only **attends to one candidate** from the other decoder



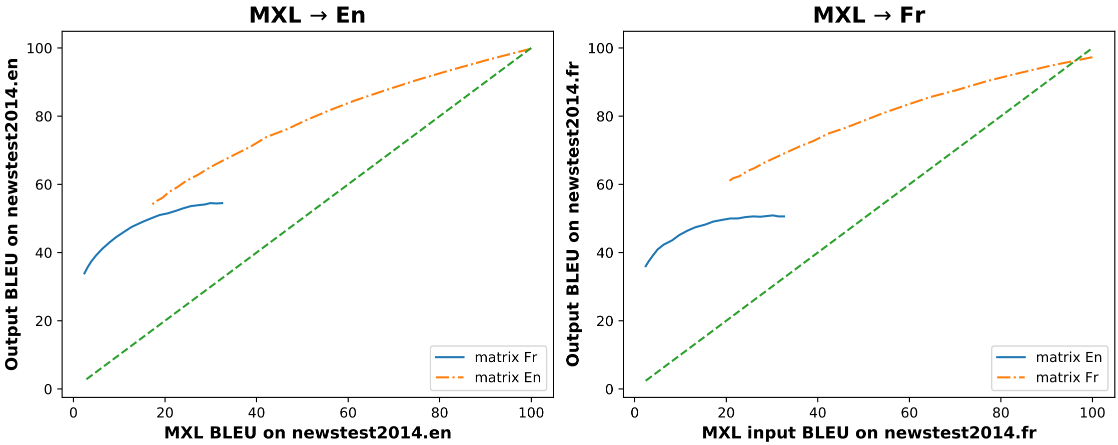
Decoding with Decoder Cross Attention

Dual beam search:

- Each candidate only **attends to one candidate** from the other decoder
- Computing overhead $(2\times)$ since **no more incremental decoding**



The Effect of Mixing Languages



- Up to 20 replacements
- Embedded segments helps translation, especially the first few segments
- Basic grammar structure helps translation

Correcting Morphological Errors

Output of `dual` model

En	In Oregon , planners are experimenting with giving drivers different choices .
Fr	Dans l' Orégon, les planificateurs tentent l' expérience en offrant aux automobilistes différents choix .
MXL	In l' Oregon , planners **tentent** l' expérience with giving **automobilistes différents** choix .
Hyp	*En l' Oregon , les planificateurs tentent l' expérience de donner aux automobilistes différents choix .*
Noisy MXL	In l' Oregon , planners **tenter** l' expérience with giving **automobilist différent** choix .
Hyp	*Dans l' Oregon , les planificateurs **peuvent tenter** l' expérience de donner **un choix différent aux automobilistes** .*

Multi-target Translation

- De→En/Fr, En→De/Fr and En→Zh/Ja
- IWSLT17 as training data (~ 200k), IWSLT TED `tst2014` as test data
- **Multilingual pre-training** with WMT data

Model	Avg[2] BLEU	Avg[2] SIM
`base`	26.7	87.53
`multi`	25.8 (-0.9)	89.05 (+1.52)
`indep`	**27.6 (+0.9)**	88.28 (+0.75)
`dual`	**26.6 (-0.1)**	88.71 (+1.18)
`indep ps`	**27.4 (+0.7)**	88.69 (+1.16)
`dual ps`	**27.3 (+0.6)**	89.00 (+1.47)
`indep FT`	**30.3 (+3.6)**	89.54 (+2.01)
`dual FT`	**30.1 (+3.4)**	89.66 (+2.13)

- **`dual`** worse than **`indep`**, possibly suffering from exposure bias problem
- Using synthetic pseudo tri-parallel data helps
- Fine-tuning using pre-trained multilingual models is beneficial

[2] Average over 3 directions: De→En/Fr, En→De/Fr and En→Zh/Ja.

Multi-target Translation

- De→En/Fr, En→De/Fr and En→Zh/Ja
- IWSLT17 as training data (~ 200k), IWSLT TED `tst2014` as test data
- **Multilingual pre-training** with WMT data

Model	Avg[2] BLEU	Avg[2] SIM
`base`	26.7	87.53
`multi`	25.8 (-0.9)	89.05 (+1.52)
`indep`	**27.6 (+0.9)**	88.28 (+0.75)
`dual`	**26.6 (-0.1)**	**88.71 (+1.18)**
`indep ps`	**27.4 (+0.7)**	88.69 (+1.16)
`dual ps`	**27.3 (+0.6)**	**89.00 (+1.47)**
`indep FT`	**30.3 (+3.6)**	89.54 (+2.01)
`dual FT`	**30.1 (+3.4)**	**89.66 (+2.13)**

- **`dual`** worse than **`indep`**, possibly suffering from exposure bias problem
- Using synthetic pseudo tri-parallel data helps
- Fine-tuning using pre-trained multilingual models is beneficial
- Higher similarity between translations

[2] Average over 3 directions: De→En/Fr, En→De/Fr and En→Zh/Ja.

Bidirectional Decoding

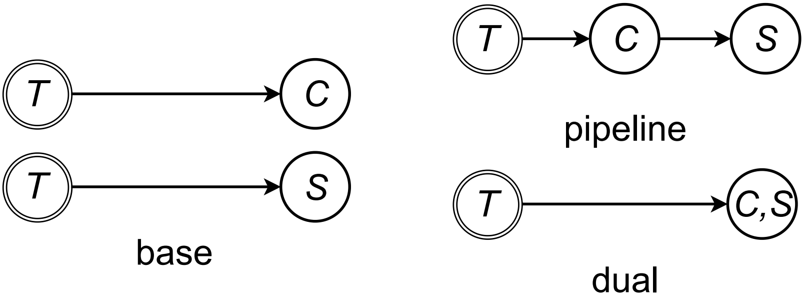
- En$\rightarrow$De, Fr, Zh, Ja
- Same data as multi-target translation

Model	Avg[3] BLEU	Avg[3] Consistency
`base`	25.7	-
`indep`	26.5 (+0.8)	52.4
`dual`	**21.8 (-3.9)**	**83.5 (+31.1)**
`indep pseudo`	26.9 (+1.2)	62.4
`dual pseudo`	**26.5 (+0.8)**	80.3 (+17.9)

- Severe exposure bias problem for **`dual`**: low BLEU score but high consistency
- Mitigated using pseudo parallel data
- More consistent translations

[3] Average over 4 directions: En$\rightarrow$De/Fr/Zh/Ja.

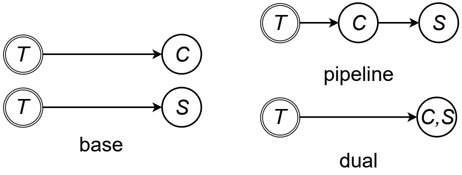
Multilingual Subtitling



- MuST-Cinema En-Fr data
- ~ 275k for training, 544 for test
- WMT data $(33.9$M$)$ for pre-training

Multilingual Subtitling

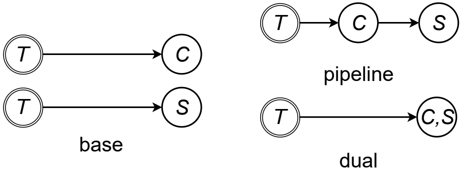
Model	BLEU			Consistency	
	EN	FR	Avg	Structural	Lexical
base	55.7	23.9	39.8	55.3	70.7
base +FT	55.7	24.9	40.3	54.5	71.4
pipeline	55.7	**23.6**	39.7	**95.7**	**96.0**
pipeline +FT	55.7	**24.2**	40.0	**98.4**	**98.3**
dual +FT	56.9	25.6	41.3	65.1	79.1
share +FT	56.5	25.8	41.2	66.7	80.0



- Pipeline **worse in quality**, **higher in consistency**

Multilingual Subtitling

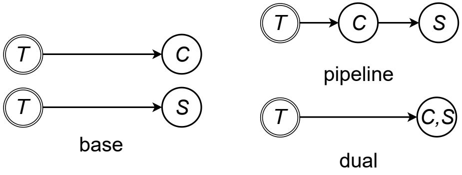
Model	BLEU			Consistency	
	EN	FR	Avg	Structural	Lexical
base	55.7	23.9	39.8	55.3	70.7
base +FT	55.7	24.9	40.3	54.5	71.4
pipeline	55.7	23.6	39.7	95.7	96.0
pipeline +FT	55.7	24.2	40.0	98.4	98.3
dual +FT	**56.9**	**25.6**	**41.3**	**65.1**	**79.1**
share +FT	56.5	25.8	41.2	66.7	80.0



- Pipeline **worse in quality**, **higher in consistency**
- `dual` **improves translation quality**, with **higher consistency** than base

Multilingual Subtitling

Model	BLEU			Consistency	
	EN	FR	Avg	Structural	Lexical
base	55.7	23.9	39.8	55.3	70.7
base +FT	55.7	24.9	40.3	54.5	71.4
pipeline	55.7	23.6	39.7	95.7	96.0
pipeline +FT	55.7	24.2	40.0	98.4	98.3
dual +FT	56.9	25.6	41.3	65.1	79.1
share +FT	56.5	25.8	41.2	**66.7**	**80.0**



- Pipeline **worse in quality**, **higher in consistency**
- **`dual` improves translation quality**, with **higher consistency** than base
- **Sharing decoder parameters** delivers **similar results**, **better consistency** than `dual`, and **fewer parameters**