廣东外語外貿大學
GUANGDONG UNIVERSITY OF FOREIGN STUDIES

IALP 2024
The 28th International Conference on Asian Language Processing, 4-6 August 2024, Hohhot, China

Cross-Lingual Word Alignment for ASEAN Languages with Contrastive Learning

Jingshen Zhang[1], Xinying Qiu*[1], Teng Shen[1],Wenyu Wang[3],
Kailin Zhang[1] and Wenhe Feng[2]

[1]*School of Information Science and Technology*
[2]*Laboratory of Language Engineering and Computing*
Guangdong University of Foreign Studies, Guangzhou, China
[3]*College of Computer and Software Engineering, Hohai University, Nanjing, China*

Reporter: Jingshen Zhang
Email: audbut0702@163.com; xy.qiu@foxmail.com

Cross-Lingual Word Alignment

TABLE I. AN EXAMPLE OF WORD ALIGNMENT

Input	**Chinese:** 我是一个生态学家，我研究复杂性 **Lao:** tôi là một nhà sinh thái học và tôi nghiên cứu sự phức tạp
Output	(tôi, 我), (là, 是), (một, 一个), (nhà sinh thái học, 生态学家), (nghiên cứu, 研究), (sự phức tạp, 复杂性)

Cross-Lingual Word Alignment

- Aiming to **identify and align word-level** correspondences between **parallel sentences** in two languages.
- It plays a critical role in various downstream applications, such as machine translation, bilingual lexicon induction and many other areas.

Challenge

However, word alignment remains challenging for **low-resource** due to **the scarcity of parallel training data**.

Previous Methodology

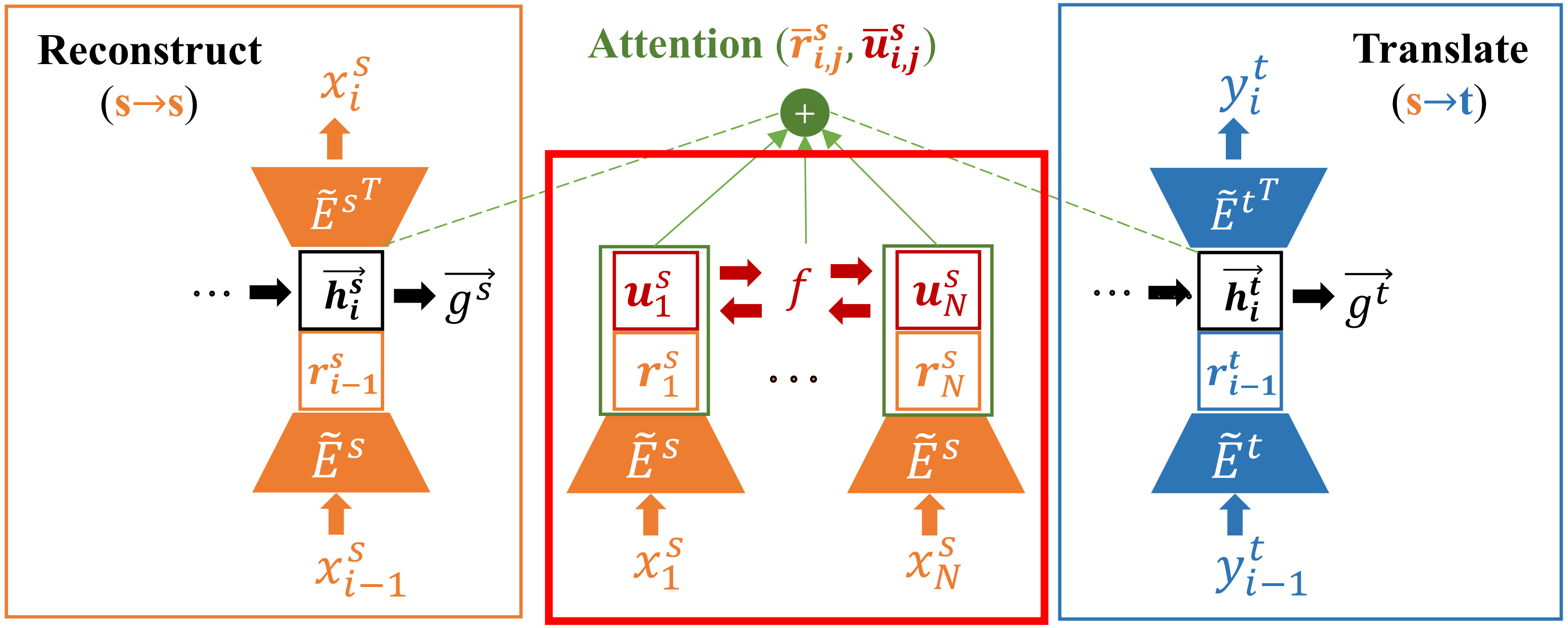


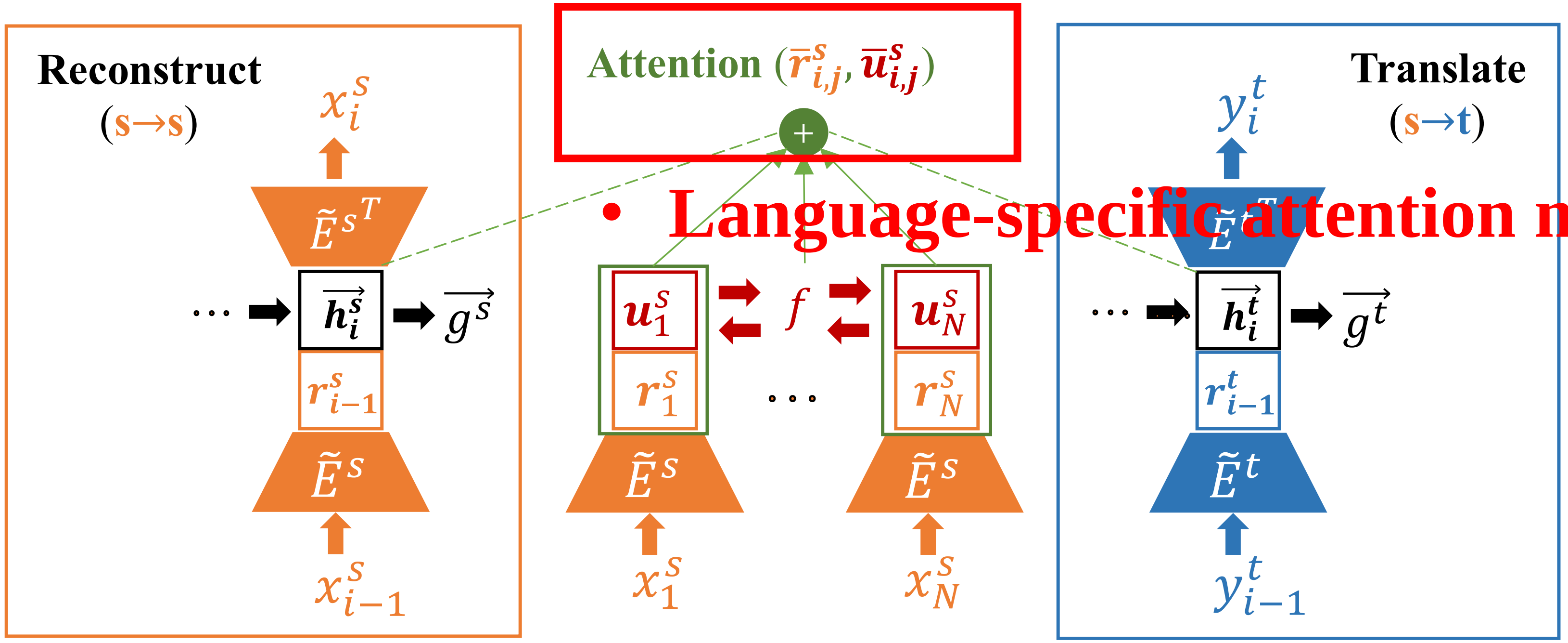
Fig 1. Model architecture proposed by Wada et al.

- **A shared bidirectional LSTM encoder**

For low-resource languages, Wada et al., (2021)[1] proposed a **BiLSTM-based encoder-decoder** model with attention:

[1] Wada et al. 2021. Learning contextualized cross-lingual word embeddings for extremely low-resource languages using parallel corpora. CORR, abs/2010.14649

Previous Methodology



- Language-specific attention mechanisms

Fig 1. Model architecture proposed by Wada et al.

For low-resource languages, Wada et al., (2021)[1] proposed a **BiLSTM-based encoder-decoder** model with attention:

[1] Wada et al. 2021. Learning contextualized cross-lingual word embeddings for extremely low-resource languages using parallel corpora. CORR, abs/2010.14649

Previous Methodology

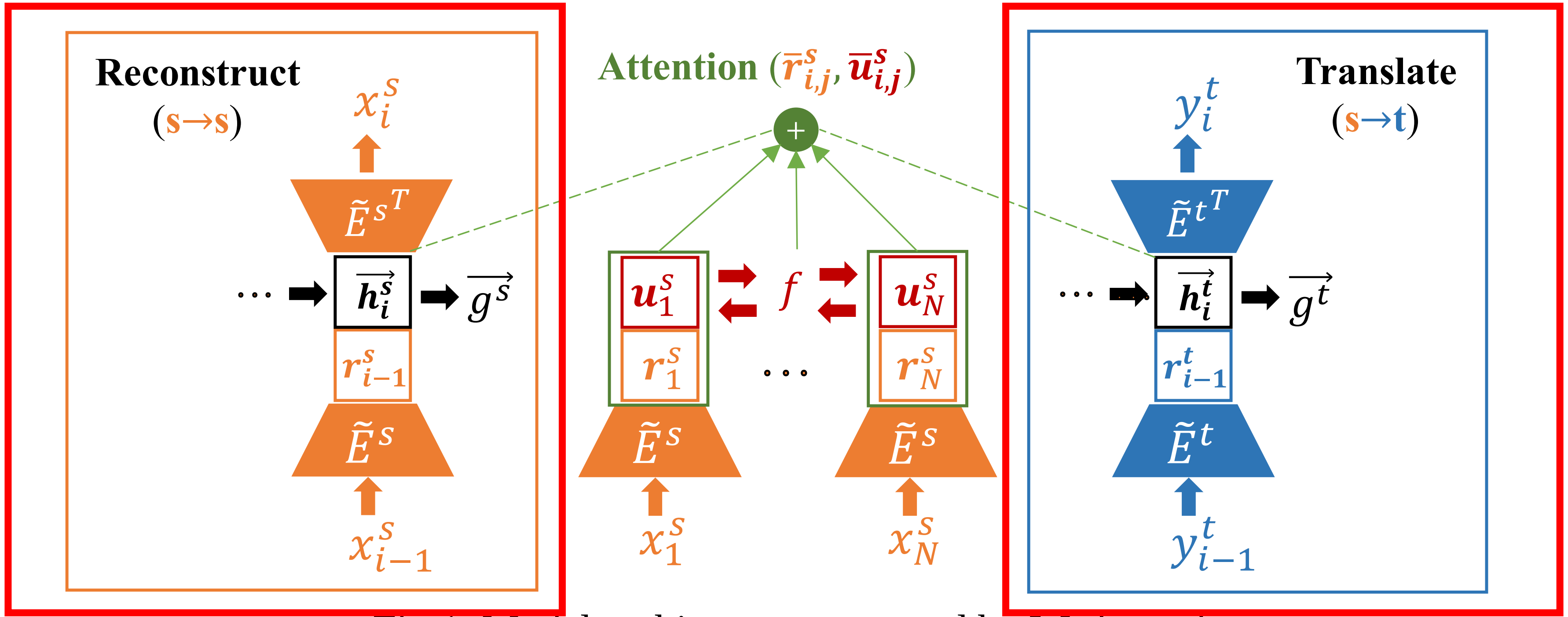


Fig 1. Model architecture proposed by Wada et al.

- **Language-specific unidirectional LSTM decoder**

For low-resource languages, Wada et al., (2021)[1] proposed a **BiLSTM-based encoder-decoder** model with attention:

[1] Wada et al. 2021. Learning contextualized cross-lingual word embeddings for extremely low-resource languages using parallel corpora. CORR, abs/2010.14649

Research Gap in Previous Methodology

- Previous approach does not **explicitly model the relationships between words** in the embeddings space.

- **Potential of contrastive learning for cross-lingual word alignment** has not been fully explored, particularly in low-resource settings.

Goal of Our Work

- Improve the performance of the BiLSTM-based encoder-decoder model by **incorporating contrastive learning**.

- Conduct extensive experiments on **4 ASEAN languages, 5 pairwise datasets.**

Our Methodology

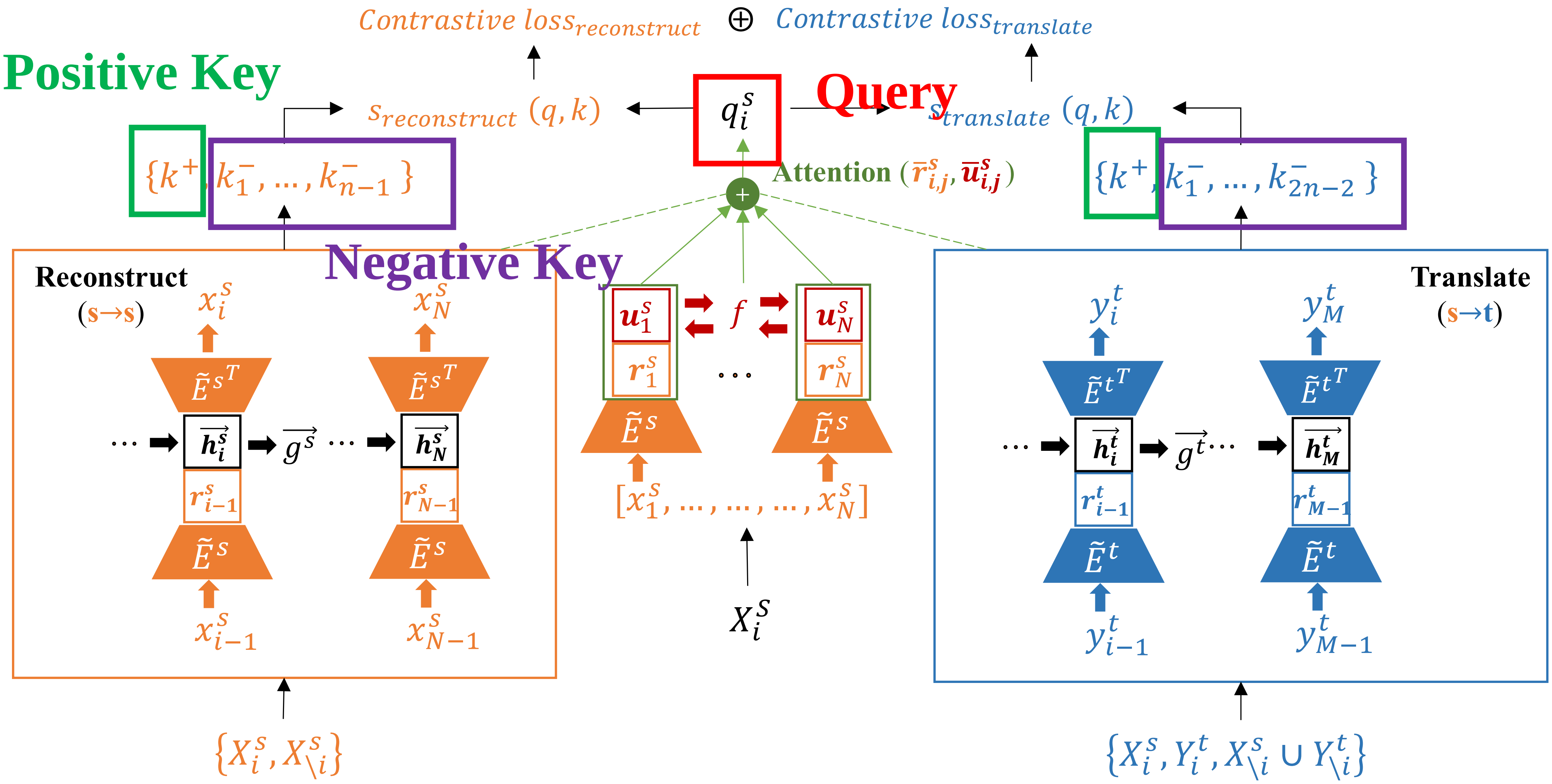


Fig 2. Our proposed model architecture

GUANGDONG UNIVERSITY OF FOREIGN STUDIES

Our Methodology

IALP 2024

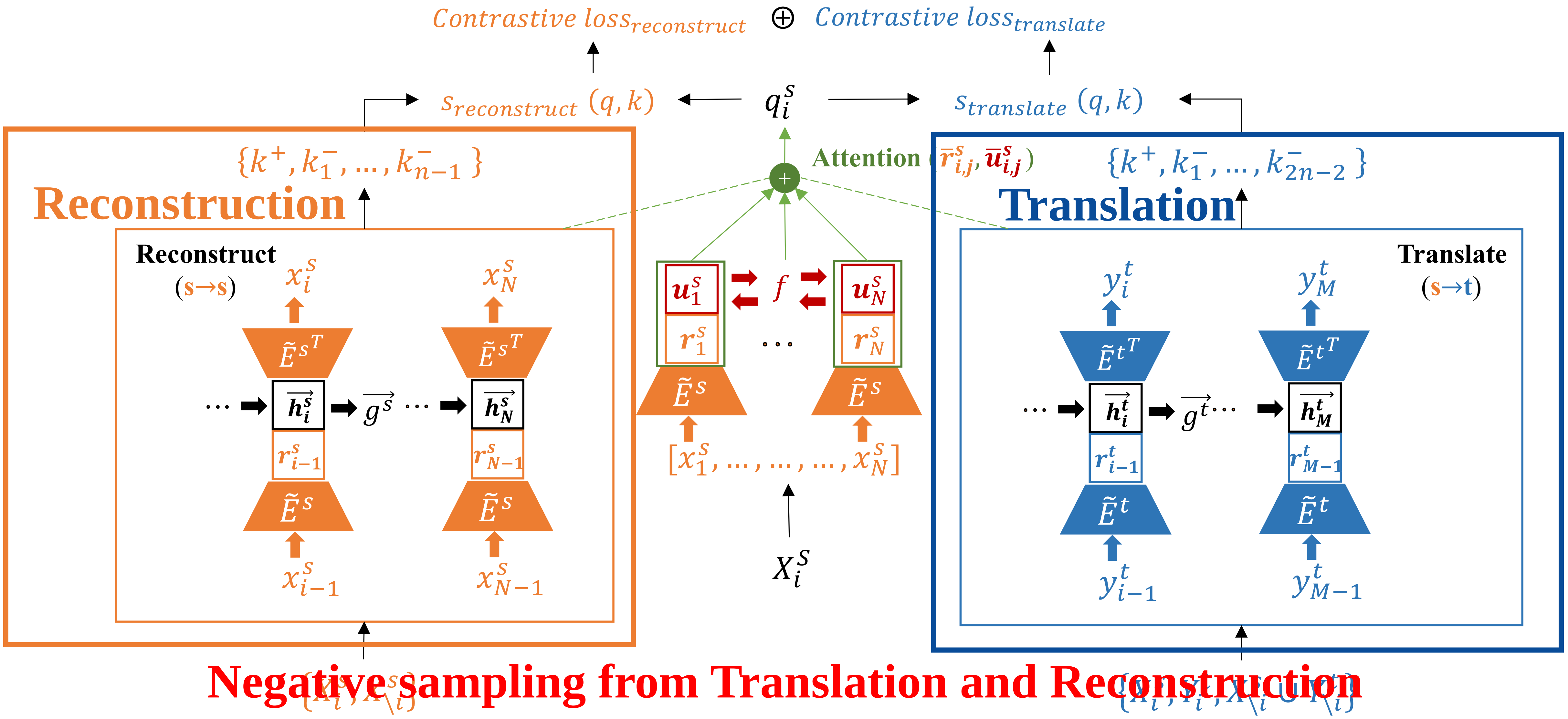


Fig 2. Our proposed model architecture

Our Methodology

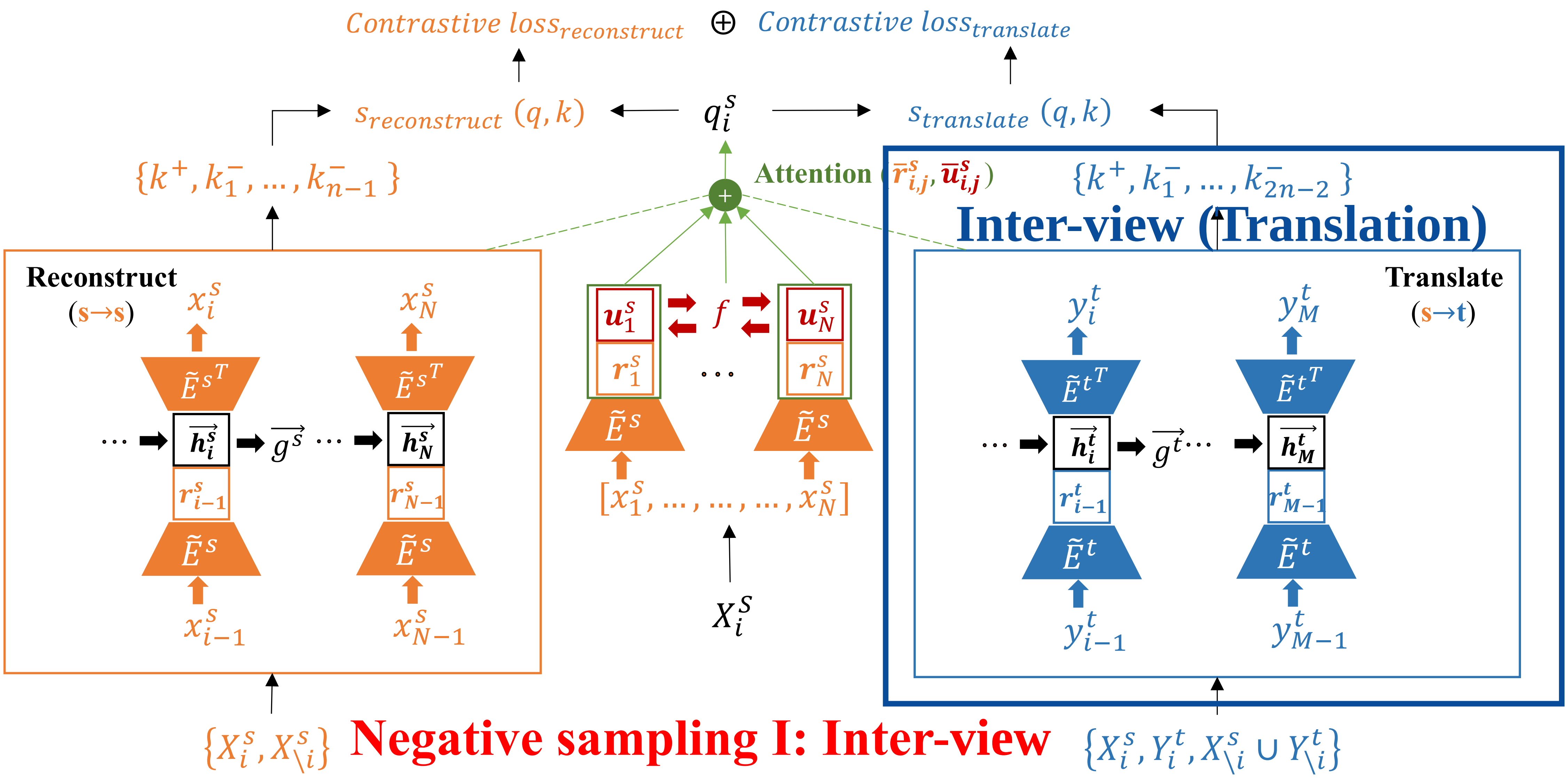


Fig 2. Our proposed model architecture

Our Methodology

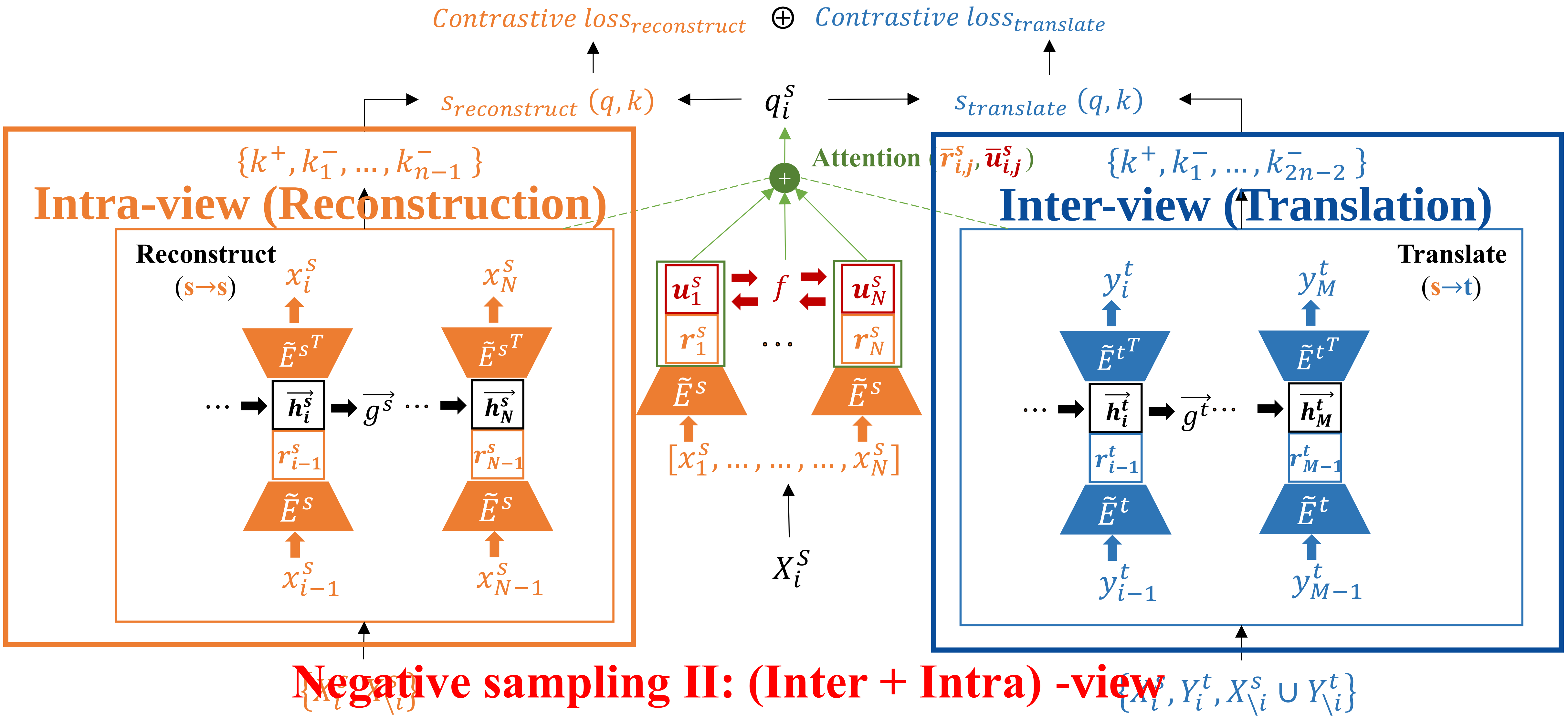


Fig 2. Our proposed model architecture

GUANGDONG UNIVERSITY OF FOREIGN STUDIES

Contrastive Loss Function Construction

Source Sentence

$$\mathcal{L}_{ctl}(X_i) = -log\frac{e^{\theta(q,k^{+})/\tau}}{e^{\theta(q,k^{+})/\tau}+logit_{inter}+logit_{intra}\cdot\mu}$$

$$logit_{inter} = \sum_{j=1}^{y\backslash i} e^{\theta(q,k_j^{-})/\tau}, \qquad logit_{intra} = \sum_{j=1}^{x\backslash i} e^{\theta(q,k_j^{-})/\tau}$$

$$\mu\begin{cases} 0, & inter - view \\ 1, & (inter + intra) - view \end{cases}$$

IALP 2024

Contrastive Loss Function Construction

Target Sentence

$$\mathcal{L}_{ctl}(Y_i) = -log\frac{e^{\theta(\tilde{q},\tilde{k}^{+})/\tau}}{e^{\theta(\tilde{q},\tilde{k}^{+})/\tau}+\widetilde{logit}_{inter}+\widetilde{logit}_{intra}\cdot\mu}$$

$$\widetilde{logit}_{inter} = \sum_{j=1}^{y\backslash i} e^{\theta(\tilde{q},\tilde{k}_j^{-})/\tau}, \qquad \widetilde{logit}_{intra} = \sum_{j=1}^{x\backslash i} e^{\theta(\tilde{q},\tilde{k}_j^{-})/\tau}$$

$$\mu\begin{cases} 0, & inter - view \\ 1, & (inter + intra) - view \end{cases}$$

Contrastive Loss Function Construction

- Contrastive Loss for **Translation**:

$$\mathcal{L}_{translation} = \frac{1}{2n}\sum_{i=1}^{n}\{\mathcal{L}_{ctl}(X_i) + \mathcal{L}_{ctl}(Y_i)\}$$

- Contrastive Loss for **Reconstruction**:

$$\mathcal{L}_{reconstruct} = \frac{1}{n}\sum_{i=1}^{n}\left\{-log\frac{e^{\theta(q,k^+)/\tau}}{e^{\theta(q,k^+)/\tau}+logit_{intra}}\right\}$$

Contrastive Loss Function Construction

- **Combined Loss** for our proposed strategy:

$$\mathcal{L}_{combine}\begin{cases}\mathcal{L}_{translation} = \frac{1}{2n}\sum_{i=1}^{n}\{\mathcal{L}_{ctl}(X_i)+\mathcal{L}_{ctl}(Y_i)\}, & s \neq t \\ \mathcal{L}_{reconstruct} = \frac{1}{n}\sum_{i=1}^{n}\{\mathcal{L}_{ctl}(X_i)\}, & s = t\end{cases}$$

Dataset

- **4 ASEAN**: Lao, Indonesian, Vietnamese, Thai
- **5 Pairwise** Language Datasets

TABLE II. DATASET STATISTICS

Source - Target	**Pairwise sentences Source (number)**	**Pairwise words Source (number)**
Lao – Zh	OPUS + Glosbe (1503)	Glosbe (16712)
Id – Zh	OPUS (15000)	Lingea (7939)
Vi – Zh	OPUS (6916)	Lingea (4272)
Th – Zh	OPUS (50000)	Lingea (6595)
Lao - Th	OPUS + Glosbe (3190)	Lingea (2272)

Baselines (Statistic-based)

- **Fast-Align (Dyer et al., 2013)**
  - A popular statistical word aligner which serves as a streamlined and an efficient reparameteri-zation of IBM Model 2
- **GIZA++ (Och and Ney, 2003)**
  - An implementation of IBM models

Baselines (Neural-based)

- **Static PLMs (Devlin et al., 2019; Conneau et al., 2020)**
  - Extracting the static embeddings for each token. We respectively select mBERT and XLM-R
- **Sim-Align (Sabet et al., 2020)**
  - A PLM-based word aligner without finetuning on any parallel data
- **Base Model (Wada et al., 2021)**
  - Follow Wada, we train a BiLSTM-based encoder-decoder model with attention mechanism using parallel sentence pairs

Evaluation Metric

- **Cross-domain similarity local scaling** (CSLS)

$$CSLS(x,y) = 2\cos(x,y) - \frac{1}{K}\sum_{y_t \in \mathcal{N}_T(x)} \cos(x, y_t) - \frac{1}{K}\sum_{x_t \in \mathcal{N}_S(y)} \cos(x_t, y)$$

Main Results

TABLE III. MAIN RESULTS

Method	**Lao-Zh**	**Id-Zh**	**Vi-Zh**	**Th-Zh**	**Lao-Th**
Static PLM (mBERT) (Devlin et al., 2019)	1.35	36.08	17.44	3.59	4.12
Static PLM (XLM-R) (Conneau et al., 2020)	16.82	35.6	19.83	30.47	29.54
Sim-Align (mBERT) (Sabet et al., 2020)	3.38	38.88	21.02	7.27	9.19
Sim-Align (XLM-R) (Sabet et al., 2020)	17.31	40.36	23.61	35.54	37.09
Fast-Align (Dyer et al., 2013)	15.4	36	18.4	33.8	33.1
GIZA++ (Och and Ney, 2003)	16.31	67.51	20.14	52.35	44.28
Base Model (Wada et al., 2021)	54.4	73.536	44.66	62.434	69.266
Ours	**56.53**	**74.114**	**45.657**	**62.573**	**69.878**

- The BiLSTM-based model **outperforms a series of baseline models**, with an **average improvement of 38.25 P@1** over the best-performing PLM-based method.

Main Results

TABLE III. MAIN RESULTS

Method	Lao-Zh	Id-Zh	Vi-Zh	Th-Zh	Lao-Th
Static PLM (mBERT) (Devlin et al., 2019)	1.35	36.08	17.44	3.59	4.12
Static PLM (XLM-R) (Conneau et al., 2020)	16.82	35.6	19.83	30.47	29.54
Sim-Align (mBERT) (Sabet et al., 2020)	3.38	38.88	21.02	7.27	9.19
Sim-Align (XLM-R) (Sabet et al., 2020)	17.31	40.36	23.61	35.54	37.09
Fast-Align (Dyer et al., 2013)	15.4	36	18.4	33.8	33.1
GIZA++ (Och and Ney, 2003)	16.31	67.51	20.14	52.35	44.28
Base Model (Wada et al., 2021)	54.4	73.536	44.66	62.434	69.266
Ours	**56.53**	**74.114**	**45.657**	**62.573**	**69.878**

- Incorporating contrastive learning loss **further improves performance by an average of 0.75 P@1, achieving state-of-the-art results** on all fives datasets.

Main Results

TABLE IV. STATISTICAL SIGNIFICANCE TESTS

Methodology	Static PLM (mBERT)	Static PLM (XLM-R)	Sim-Align (mBERT)	Sim-Align (XLM-R)	Fast-Align	GIZA++	Base Model
T-test	***0.002**	***0.0**	*0.003	***0.0**	***0.0**	***0.023**	***0.058**

- Our model achieves **significant improvement**.

GUANGDONG UNIVERSITY OF FOREIGN STUDIES

Ablation Studies

IALP 2024

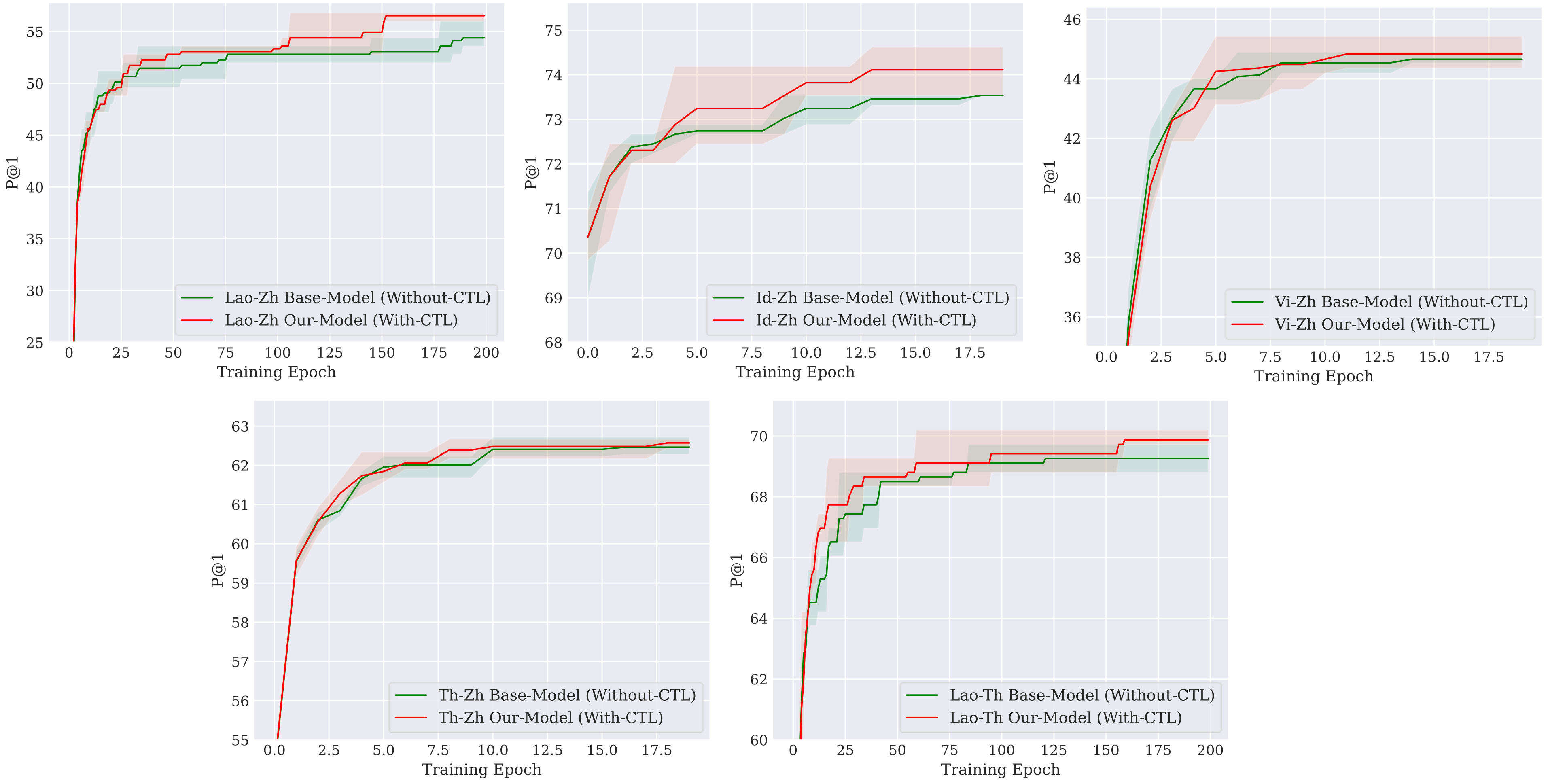


Fig 3. Compare with base model performance. The score trends over three runs with different random seed.

- Our model with contrastive learning exhibits **greater score variation**.

Ablation Studies

TABLE IV. ABLATION STUDY RESULTS

Language	**Base Model**	**Incorporating Contrastive Learning**			
		Inter (view)		***Inter + Intra (view)***	
		Avg.P.	***Max.P.***	***Avg.P.***	***Max.P.***
Lao-Zh	54.4	**56.53**↑	53.87↓	54.933↑	55.47↑
Id-Zh	73.536	73.753↑	74.042↑	**74.114**↑	73.536
Vi-Zh	44.66	45.188↑	45.129↑	**45.657**↑	45.012↑
Th-Zh	62.434	62.3↓	**62.573**↑	62.336↓	62.009↓
Lao-Th	69.286	68.96↓	69.725↑	**69.878**↑	**69.878**↑

- **Compare with Base Model**: In 4 out of 5 language pairs, integrating contrastive learning improves the base model's performance in at least 3 out of 4 metrics.

Ablation Studies

TABLE IV. ABLATION STUDY RESULTS

Language	Base Model	Incorporating Contrastive Learning			
		Inter (view)		*Inter + Intra (view)*	
		Avg.P.	***Max.P.***	***Avg.P.***	***Max.P.***
Lao-Zh	54.4	**56.53**↑	53.87↓	54.933↑	55.47↑
Id-Zh	73.536	73.753↑	74.042↑	**74.114**↑	73.536
Vi-Zh	44.66	45.188↑	45.129↑	**45.657**↑	45.012↑
Th-Zh	62.434	62.3↓	**62.573**↑	62.336↓	62.009↓
Lao-Th	69.286	68.96↓	69.725↑	**69.878**↑	**69.878**↑

- **Aggregate Function:** Average pooling slightly outperforms max pooling overall, achieving the highest scores on 4 out of 5 language pairs.

Ablation Studies

TABLE IV. ABLATION STUDY RESULTS

Language	**Base Model**	**Incorporating Contrastive Learning**			
		Inter (view)		***Inter + Intra (view)***	
		Avg.P.	***Max.P.***	***Avg.P.***	***Max.P.***
Lao-Zh	54.4	**56.53**↑	53.87↓	54.933↑	55.47↑
Id-Zh	73.536	73.753↑	74.042↑	**74.114**↑☆	73.536
Vi-Zh	44.66	45.188↑	45.129↑	**45.657**↑☆	45.012↑
Th-Zh	62.434	62.3↓	**62.573**↑	62.336↓	62.009↓
Lao-Th	69.286	68.96↓	69.725↑	**69.878**↑☆	**69.878**↑

- **Negative Sampling Strategy**: The (inter + intra)-view strategy proves to be more effective.

Conclusion

- We propose a novel approach that **incorporates contrastive learning** into a state-of-the-art BiLSTM-based encoder-decoder model for cross-lingual word alignment.
- We conduct extensive experiments on **five low-resource ASEAN** language pairs, achieving an **significant average gain of** over the base model and several strong baselines.

廣東外語外貿大學
GUANGDONG UNIVERSITY OF FOREIGN STUDIES

IALP 2024
The 28th International Conference on Asian Language Processing, 4-6 August 2024, Hohhot, China

Thank you !

明德尚行 學貫中西

Reporter: Jingshen Zhang
Email: audbut0702@163.com; xy.qiu@foxmail.com

廣东外語外貿大學
GUANGDONG UNIVERSITY OF FOREIGN STUDIES

IALP 2024
The 28th International Conference on Asian Language Processing, 4-6 August 2024, Hohhot, China

Q&A

明德尚行 學貫中西

Reporter: Jingshen Zhang
Email: audbut0702@163.com; xy.qiu@foxmail.com

Differ from previous methods

- Our approach incorporates contrastive learning into a BiLSTM-based encoder-decoder model. This allows us to explicitly model the relationships between word pairs in the cross-lingual embedding space, which previous methods did not do.
- We use positive and negative sampling strategies to refine the alignment of words across languages.

Why focus on ASEAN

- ASEAN languages are spoken by over 250 million people but have limited parallel corpora and NLP resources available. This makes them an ideal test case for low-resource cross-lingual word alignment techniques.

difference inter-view and (inter + intra) -view

- The inter-view strategy uses non-corresponding translations in the same training batch as negative instances. The inter+intra-view strategy additionally includes samples from both the source and target languages, providing more diverse negative examples.

Model perform compared PLMs

- Our model significantly outperforms pre-trained language models on these low-resource ASEAN languages. For example, on Lao-Zh, our model achieves 56.53 P@1, while mBERT and XLM-R achieve only 1.35 and 16.82 P@1 respectively.

Why contrastive learning helpful

- Contrastive learning helps the model learn a more discriminative cross-lingual embedding space by explicitly contrasting positive word pairs (translations) against negative pairs. This encourages the model to map words with similar meanings closer together while separating words with different meanings.

How deal with extremely limited data

- For these language pairs, we supplemented the OPUS corpus with additional bilingual alignment sentences from Glosbe to increase the available training data.

Why using avg pooling instead of [CLS]

- Unlike BERT-style models which use the [CLS] token, BiLSTM models typically use pooling strategies. We found that average pooling slightly outperformed max pooling in our experiments, achieving the highest scores on 4 out of 5 language pairs.

How handle subword information

- Following previous work, we use SentencePiece for subword segmentation for all languages except Chinese. For Chinese, due to its diversity, we segment words uniformly at the character level.

Why model with contrastive greater variation

- We hypothesize that this is due to differences in similarity between the randomly sampled negative and positive pairs across runs. The performance can be improved when negatives are similar but not identical to positives, as the model learns more nuanced differences. However, highly dissimilar negatives may hinder the learning process.

Extend this work to other low-resource langs

- While we haven't explicitly mentioned plans in the paper, this approach could certainly be applied to other low-resource languages. The success on ASEAN languages suggests it could be beneficial for other language families with limited resources.