Karlsruhe Institute of Technology

Modularizing NMT Systems by Standardizing Neural Representations

Jan Niehues – 22.11.2022

Motivation

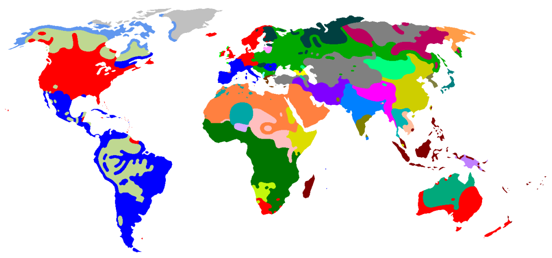
- NMT reach very good quality
  - Condition
    - Large amount of training data

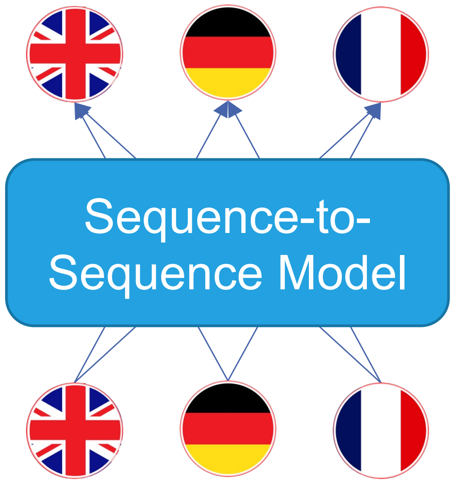
- Real-world applications
  - No End-to-End Training data
    - Parallel data between distances languages
    - Speech Translation

Multi-lingual Machine Translation

- 6000-7000 languages in the world
- Mainly focus on top 10 languages
- Minimize:
  - Human effort
  - Necessary training data

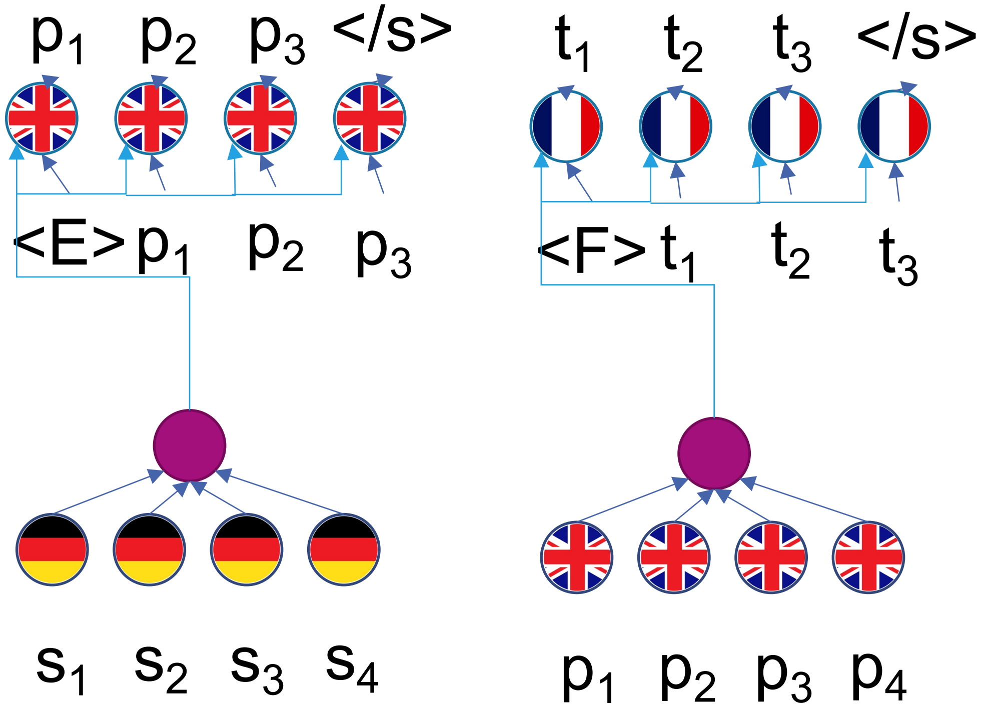
- One model to translate between many/all languages
  - Share common knowledge
  - Increase efficiency





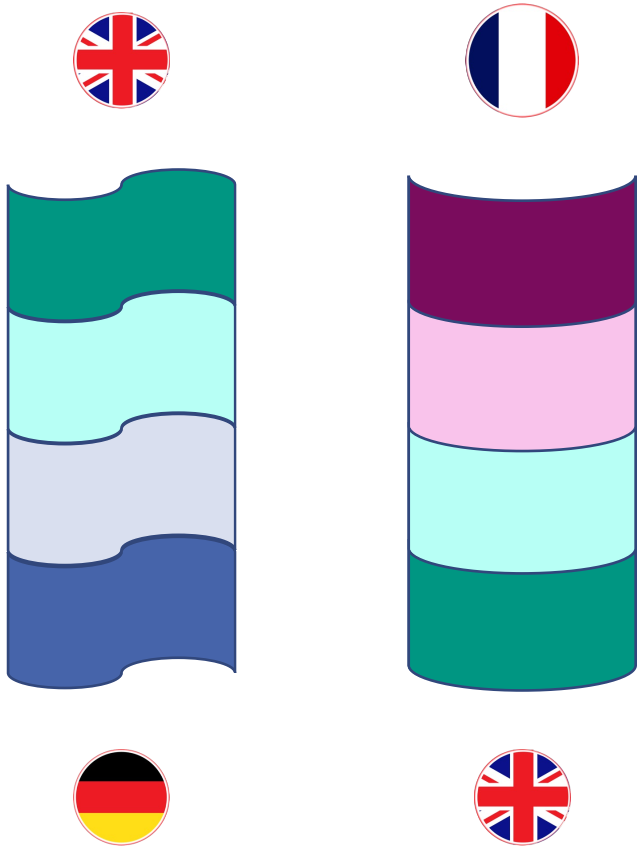
Multi-lingual Machine Translation

- One Model
  - Train on several directions
  - Control target language by <BOS>
- Challenge:
  - Generalize to unseen directions



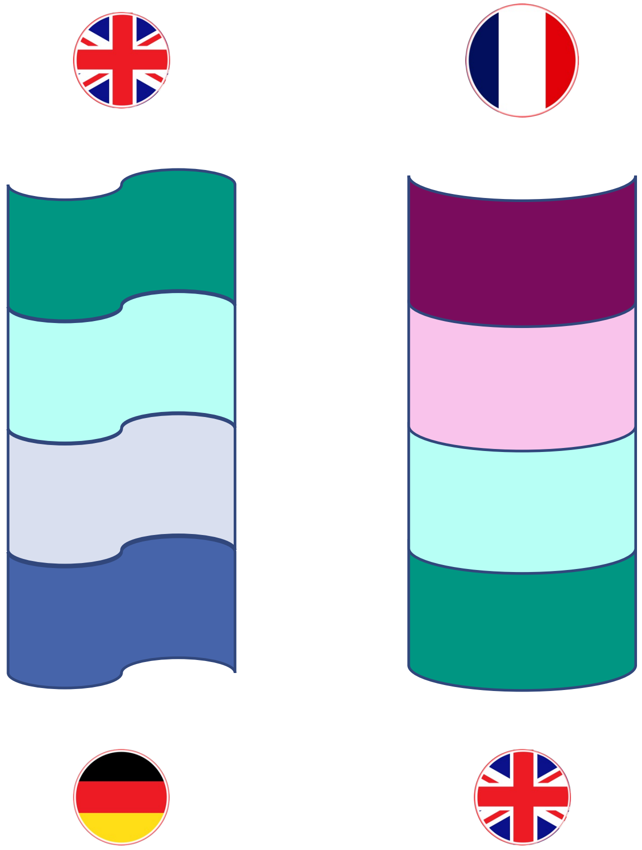
Modular Network

- Supervised directions

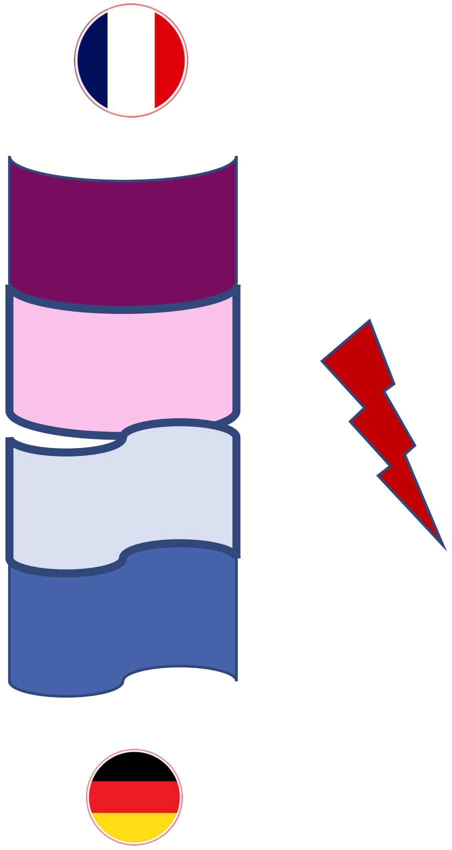


Modular Network

- Supervised directions

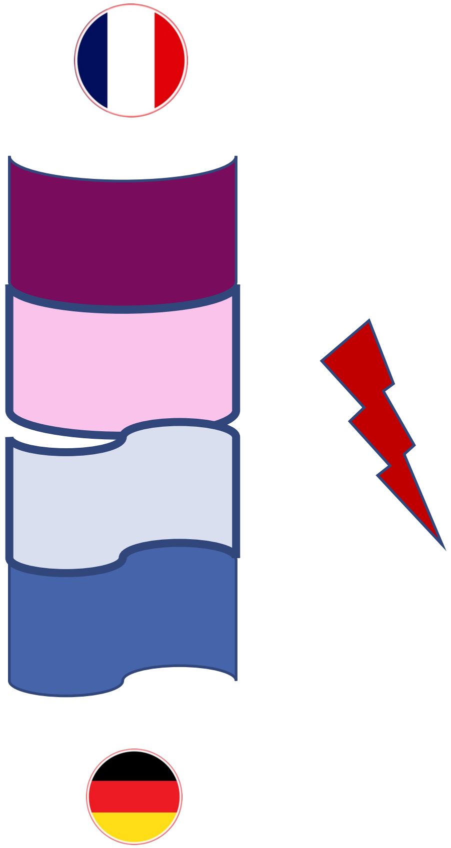


- Zero-shot directions

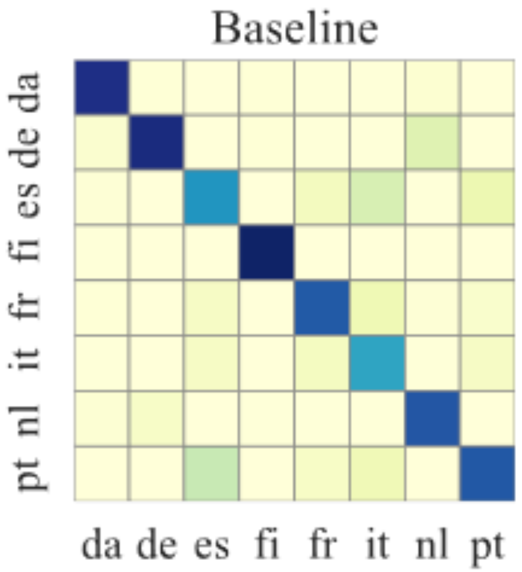


Modular Network

- Zero-shot directions

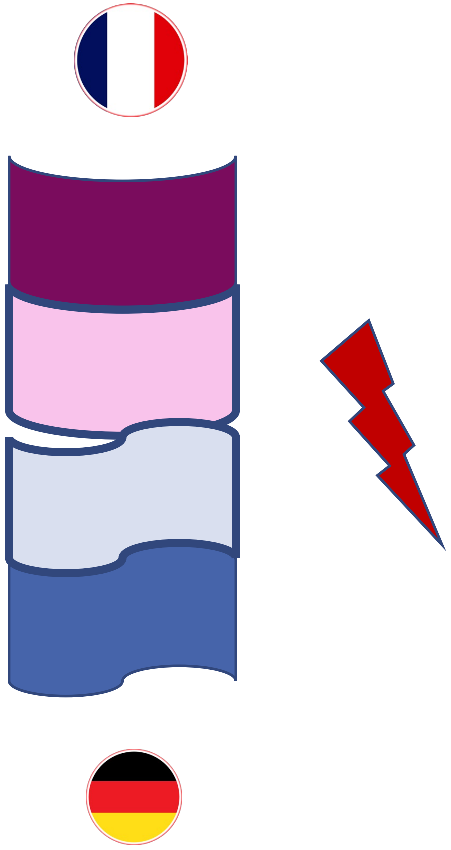


- How different are the representation?
  - How easy can we classify the source language?

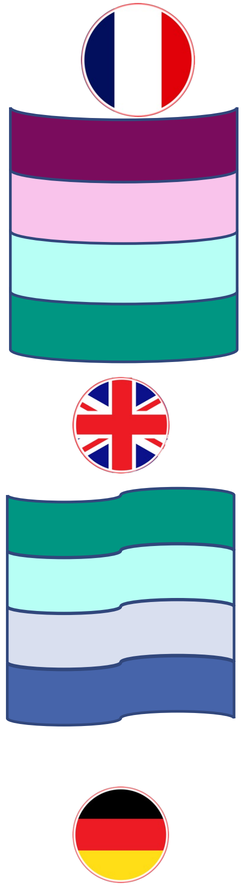


Standardizing Neural Representations

- Zero-shot directions

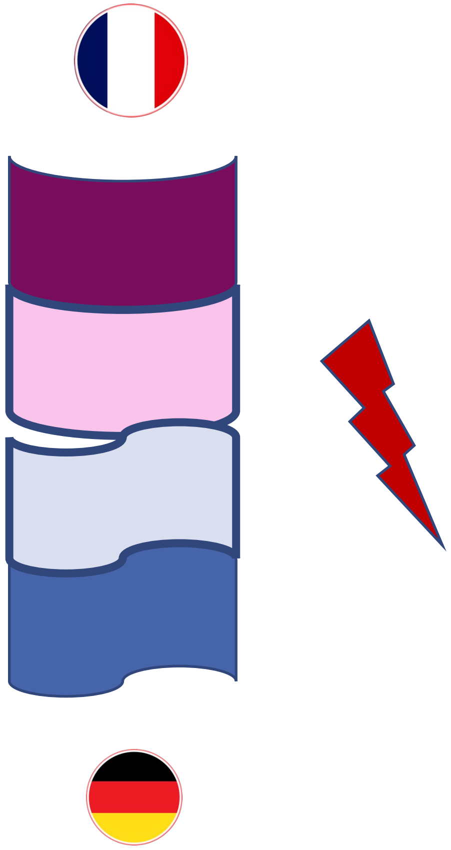


- Pivot-based translation

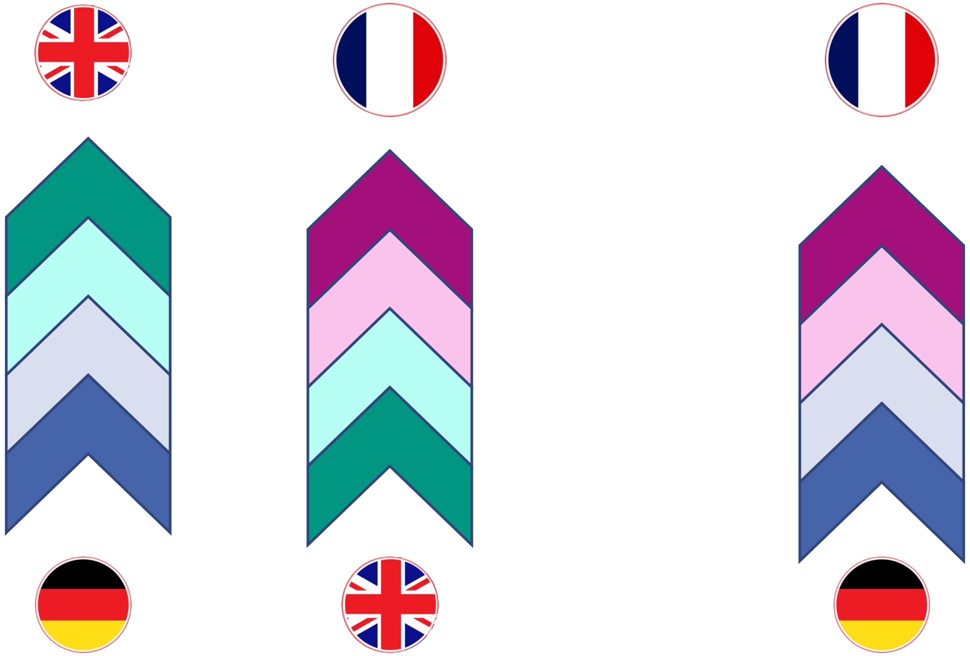


Standardizing Neural Representations

- Zero-shot directions

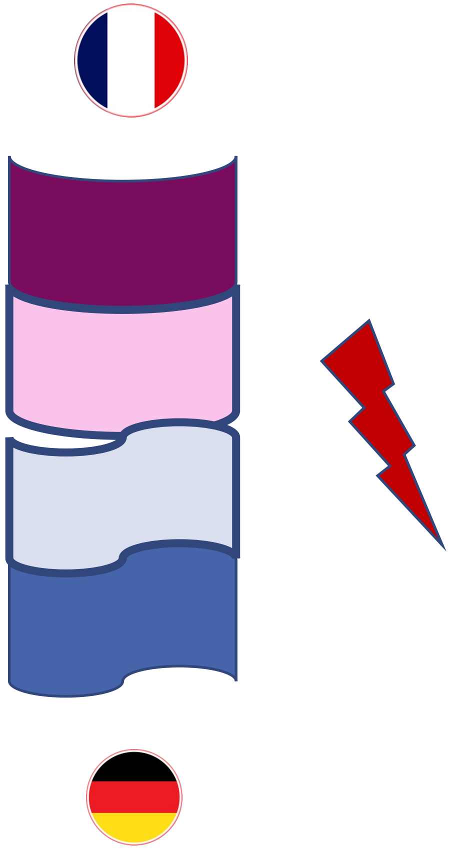


- Pivot-based translation
- Language agnostic representation
  - Continuous

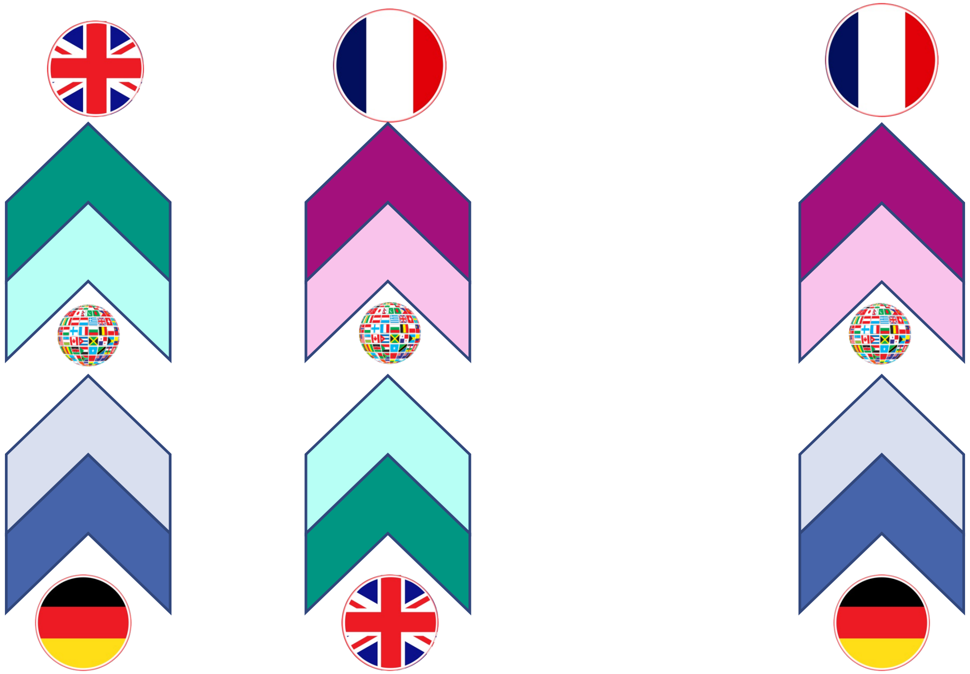


Standardizing Neural Representations

- Zero-shot directions

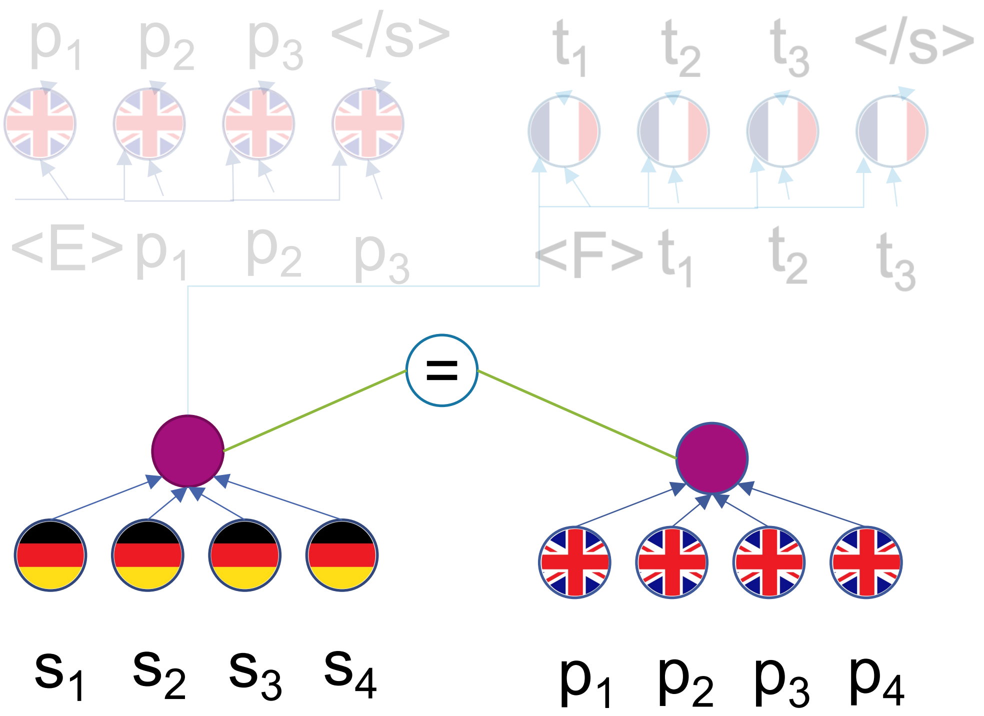


- Pivot-based translation
- Language agnostic representation
  - Continuous
  - Discrete



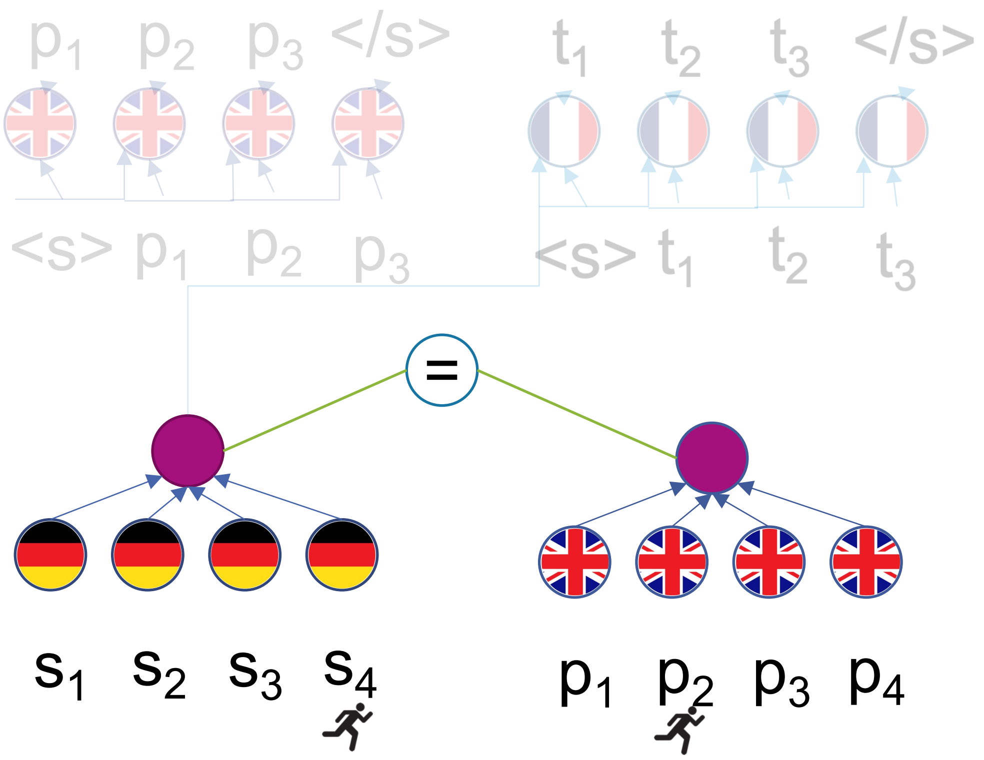
Standardizing Neural Representations

- Aim
  - Similar representation for different languages



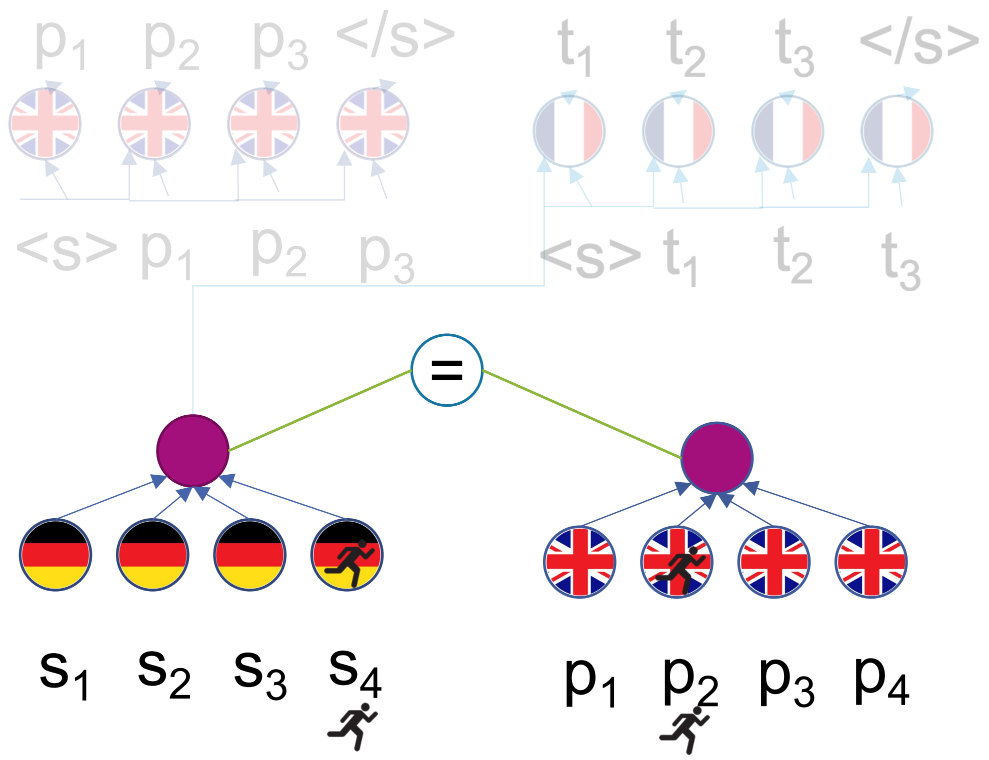
Standardizing Neural Representations

- Aim
  - Similar representation for different languages
- Challenges
  - Different word order



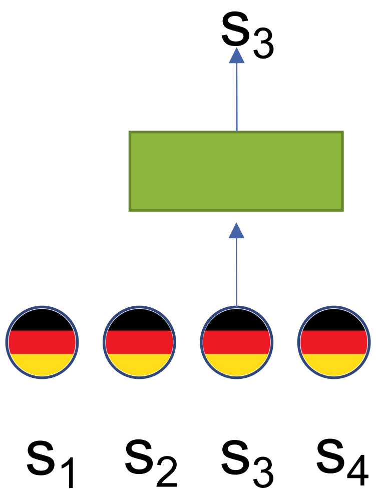
Standardizing Neural Representations

- Aim
  - Similar representation for different languages
- Challenges
  - Different word order
  - Baseline
    - 1-to-1 correspondence between words and hidden states



Analyse

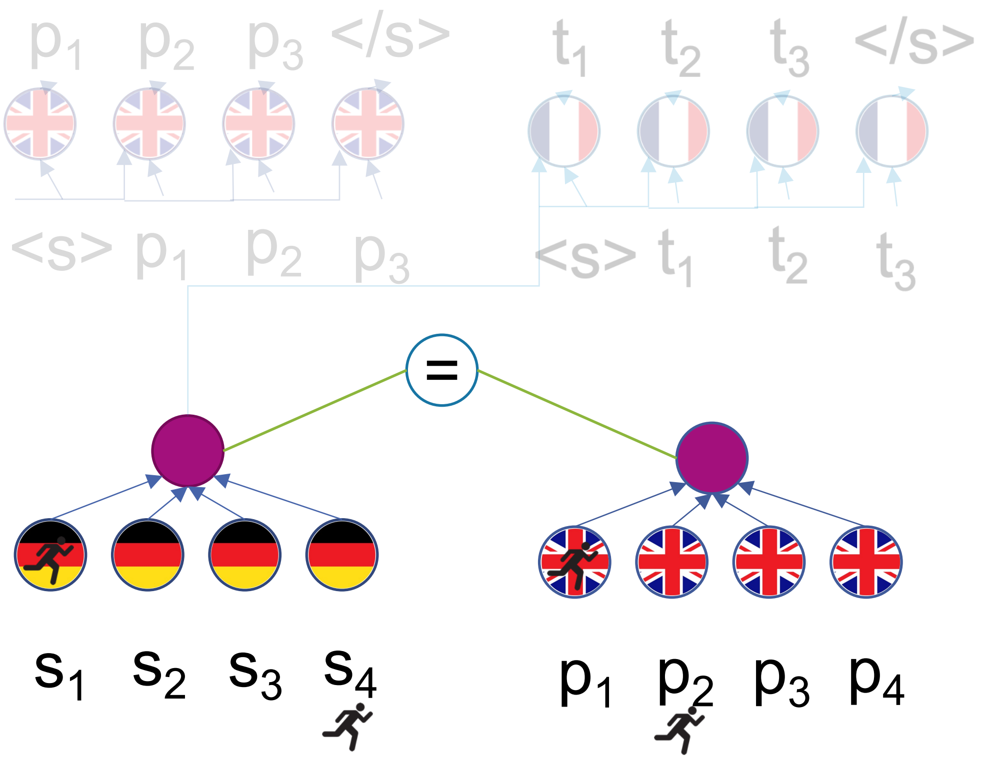
- Focus on current word
  - Transfer Learning
    - Reconstruct source word/position



Dataset	Word	Position
Baseline	99.9%	93.3%

Standardizing Neural Representations

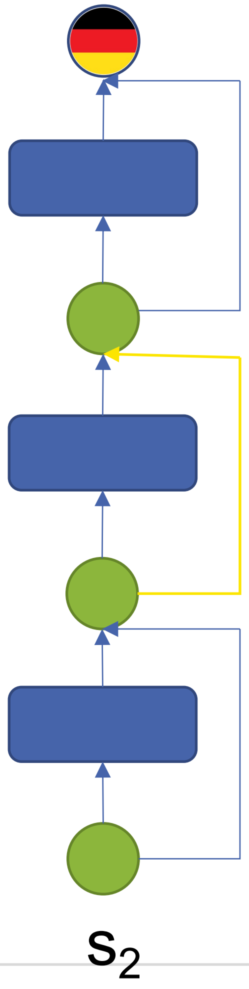
- Aim
  - Similar representation for different languages
- Idea:
  - Disentangling Positional Information



Disentangling Positional Information

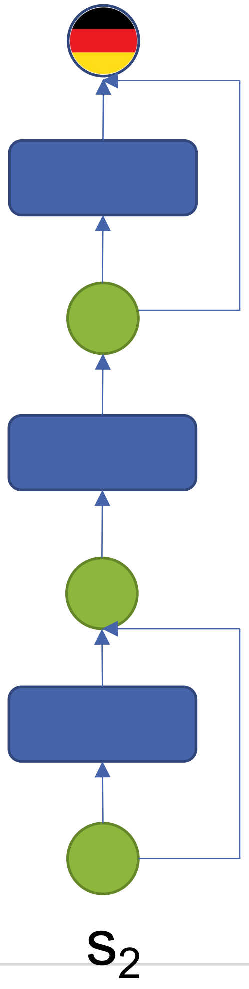
- Residual Connections
  - Shortcut
  - Improve learning

- Problem
  - Bias towards 1-to-1 correspondence between states and tokens



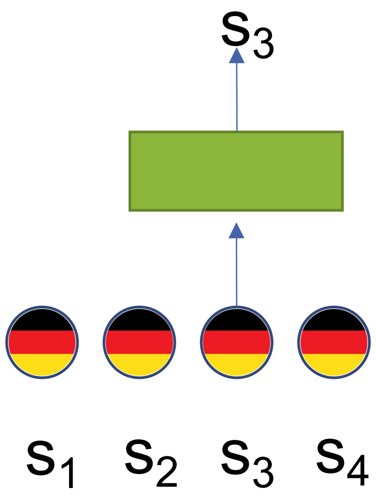
Disentangling Positional Information

- Residual Connections
  - Shortcut
  - Improve learning
- Idea:
  - Remove connection in the middle
    - Liu et al., 2021



Analyse

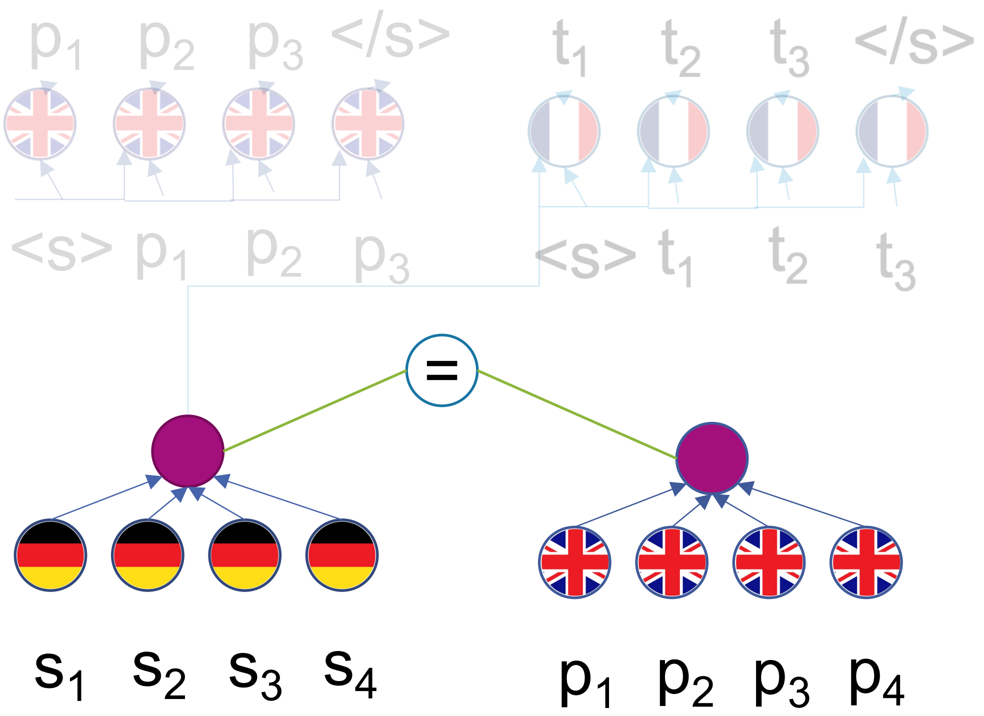
- Focus on current word
  - Transfer Learning
    - Reconstruct source word/position



Dataset	Word	Position
Baseline	99.9%	93.3%
Liu at al.	48.5%	51.4%

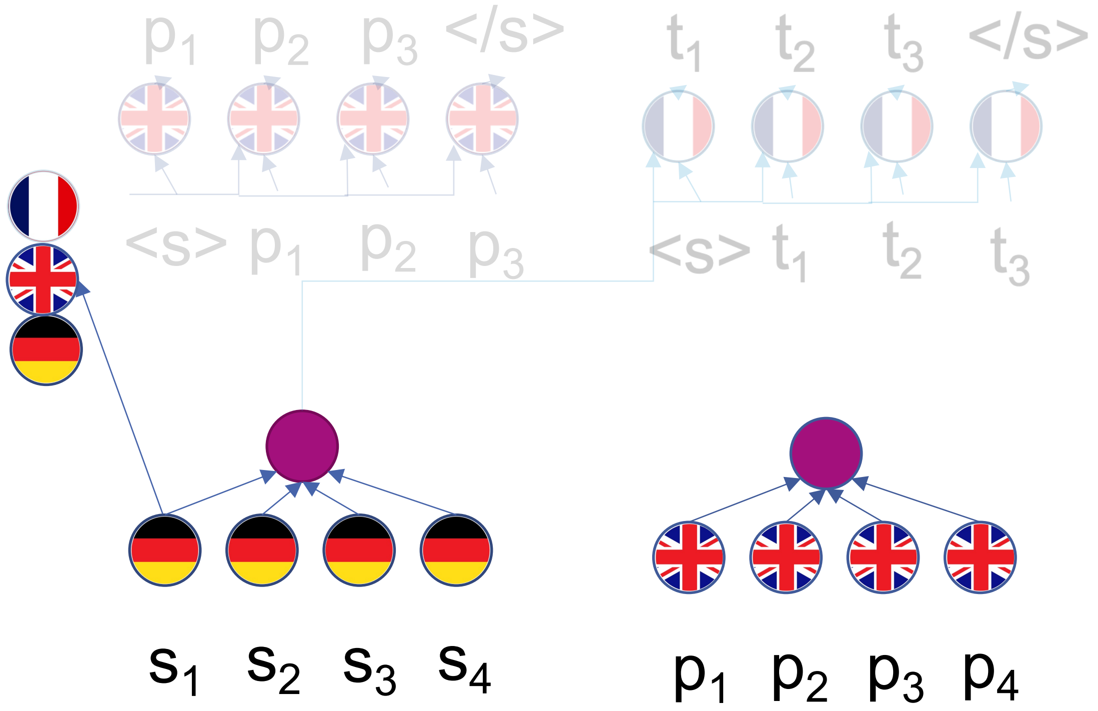
Standardizing Neural Representations

- Aim
  - Similar representation for different languages
- Idea:
  - Disentangling Positional Information
  - Similarity regularizer
    - $L_{sim} = dist\big(Encoder(x), Endocer(Y)\big)$
    - Euclidian distance between mean-pooled sentence representations
      - Arivazhagan et al. (2019)
      - Pham et al. (2019)



Standardizing Neural Representations

- Aim
  - Similar representation for different languages
- Idea:
  - Disentangling Positional Information
  - Similarity regularizer
  - Adversarial Language Classifier
    - $L_{adv} = \sum_{c=1}^{L} y_c \log(1 - p_c)$
      - Motivated by Arivazhagan et al. (2019)



Experiment

- Zero-shot translation quality
  - 3 data sets
    - Parallel data between English und 3,8 or 9 languages
  - BLEU Score

Dataset	Baseline	Disent.	Sim	Adv	Adv.+Disent
IWSLT	10.9	17.9	16.7	16.8	18.0
Europarl	13.4	25.2	24.5	25.3	26.1
PMIndia	2.4	14.3	8.9	7.3	17.1

Experiment

- Zero-shot translation quality
  - 3 data sets
    - Parallel data between English und 3,8 or 9 languages
  - BLEU Score

Dataset	Baseline	Disent.	Sim	Adv	Adv.+Disent	Pivot
IWSLT	10.9	17.9	16.7	16.8	18.0	19.1
Europarl	13.4	25.2	24.5	25.3	26.1	26.0
PMIndia	2.4	14.3	8.9	7.3	17.1	22.1

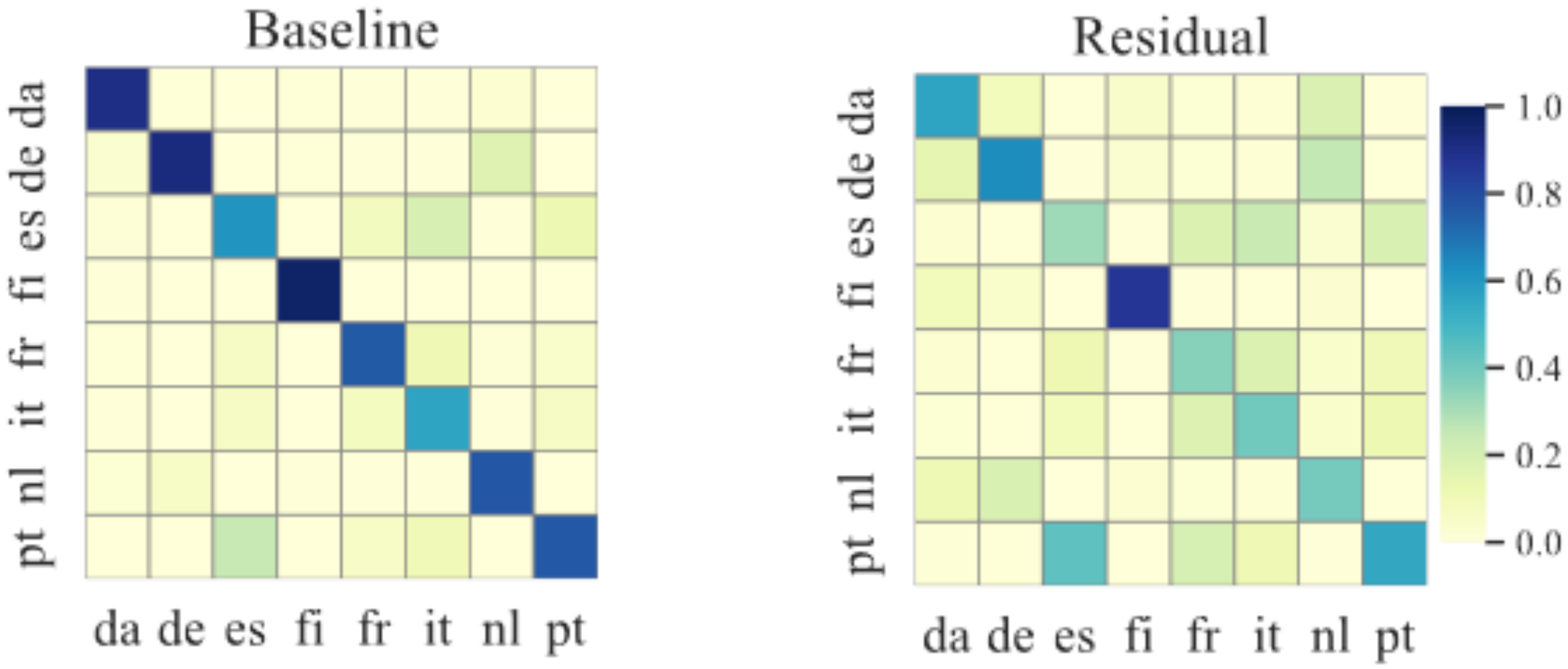
Experiment

- Related languages
  - Europal without overlapping sentences

Dataset	Baseline	Disent.	Pivot
All	8.2	26.7	27.1
Germanic	11.8	25.5	24.8
Romance	13.5	32.2	31.0

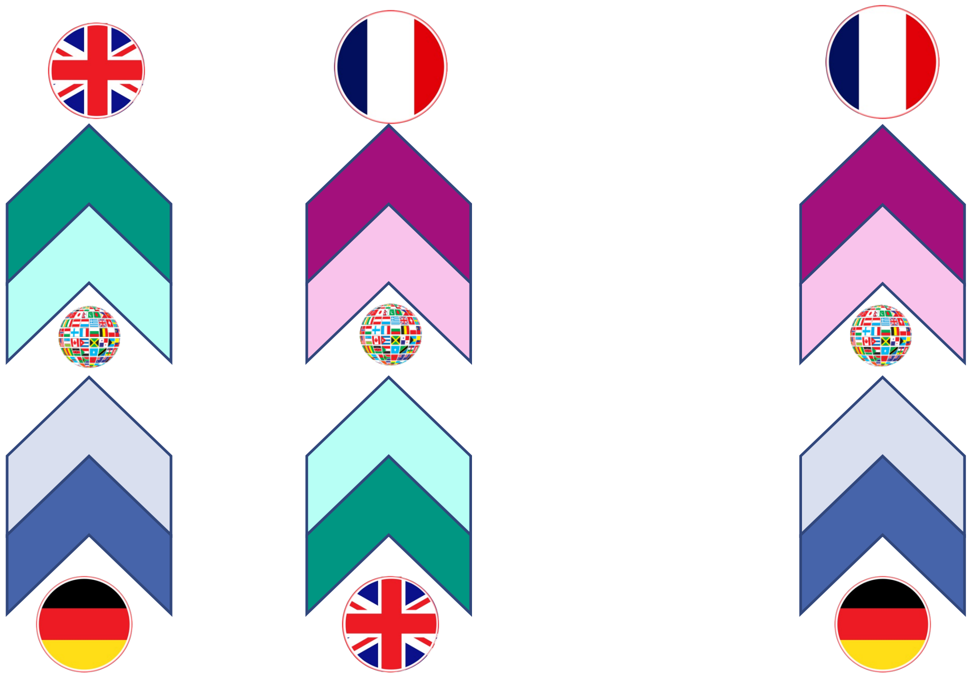
Similarity of the representations

- Classify source language of the encoder states



Discrete Representations

- Motivation
  - Construct artificial languages



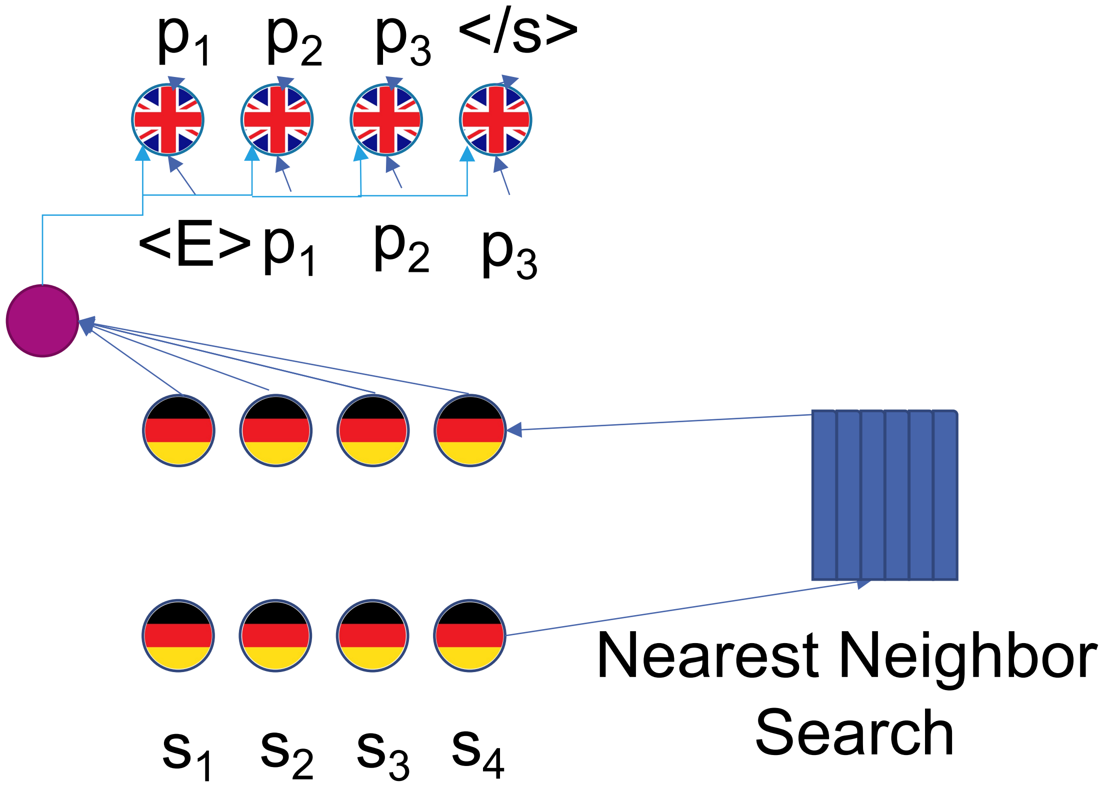
- Advantages:
  - Discrete representation are more robust
  - Interpretation
- Example

source sentence (English)	learning	a	new	language
	↓	↓	↓	↓
discrete codes	3	609	57	1042

source sentence (Indonesian)	belajar	bahasa	baru
	↓	↓	↓
discrete codes	3	57	258

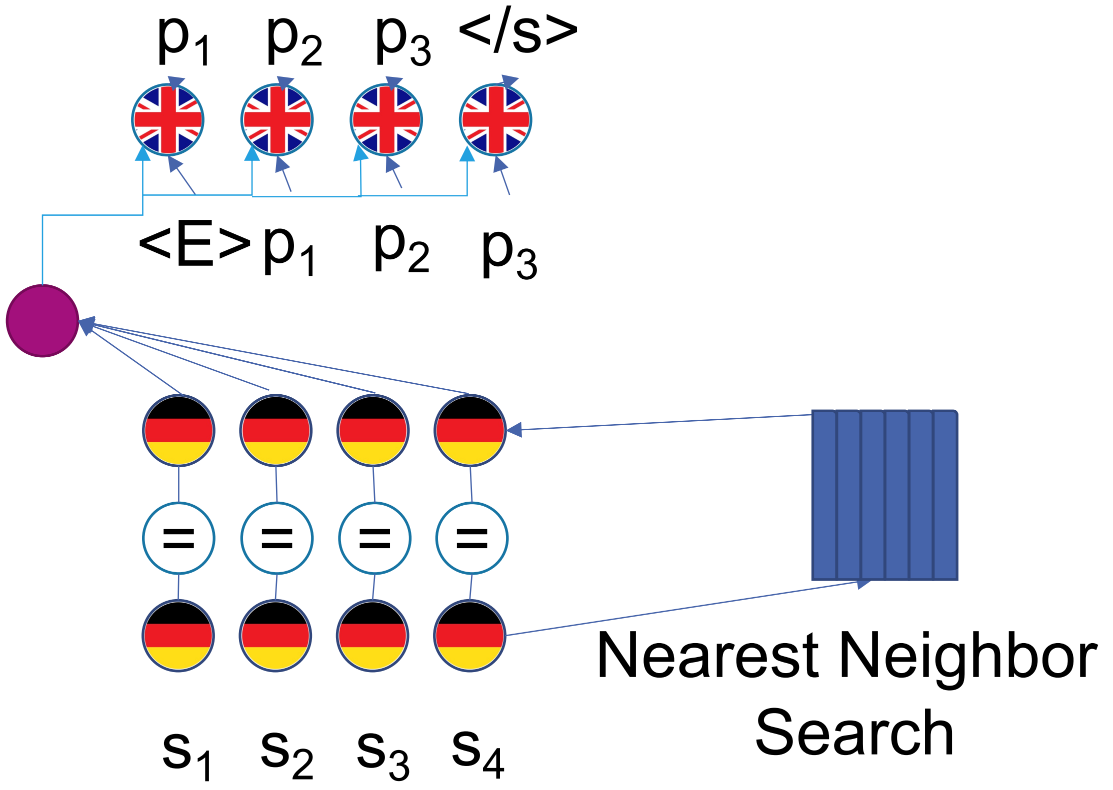
Discrete Representations

- Challenge:
  - Learning representation
    - Codebook



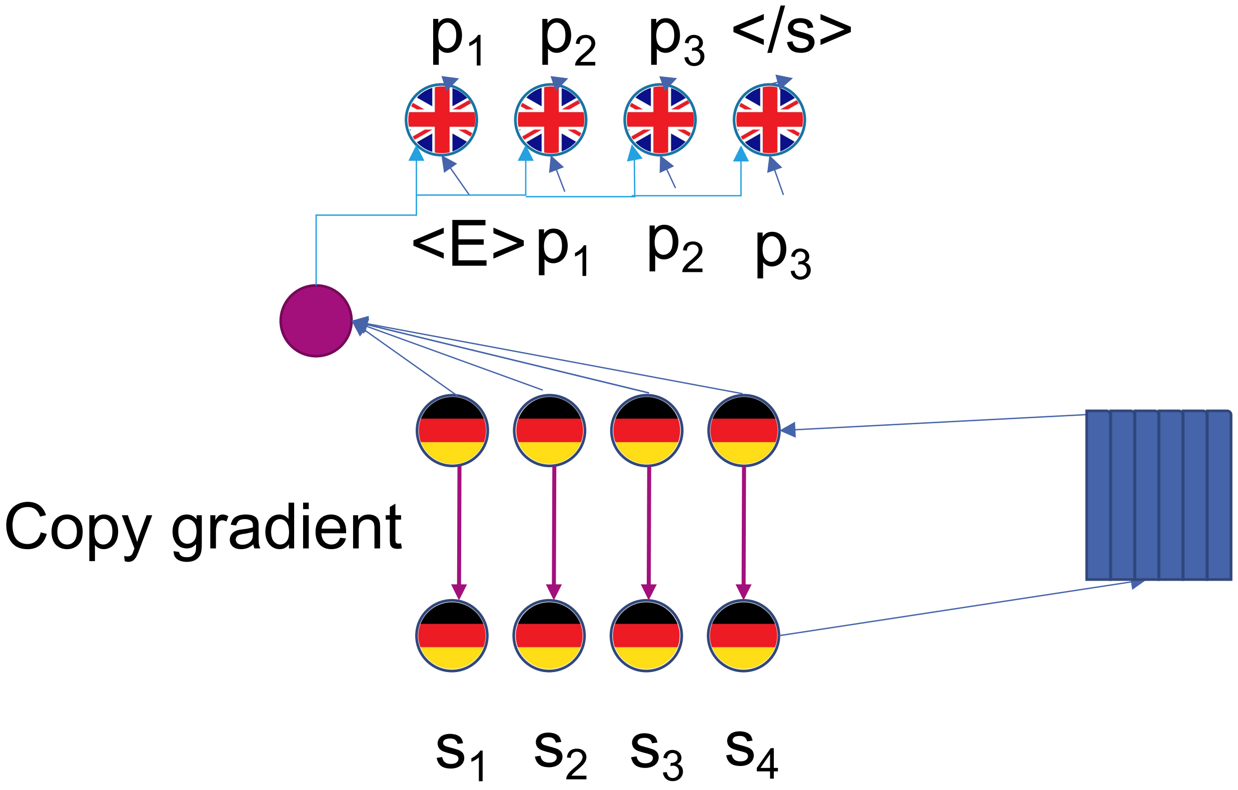
Discrete Representations

- Challenge:
  - Learning representation
    - Codebook
    - Minimize discretization error
      - $L = |enc(X) - q(enc(x))|$



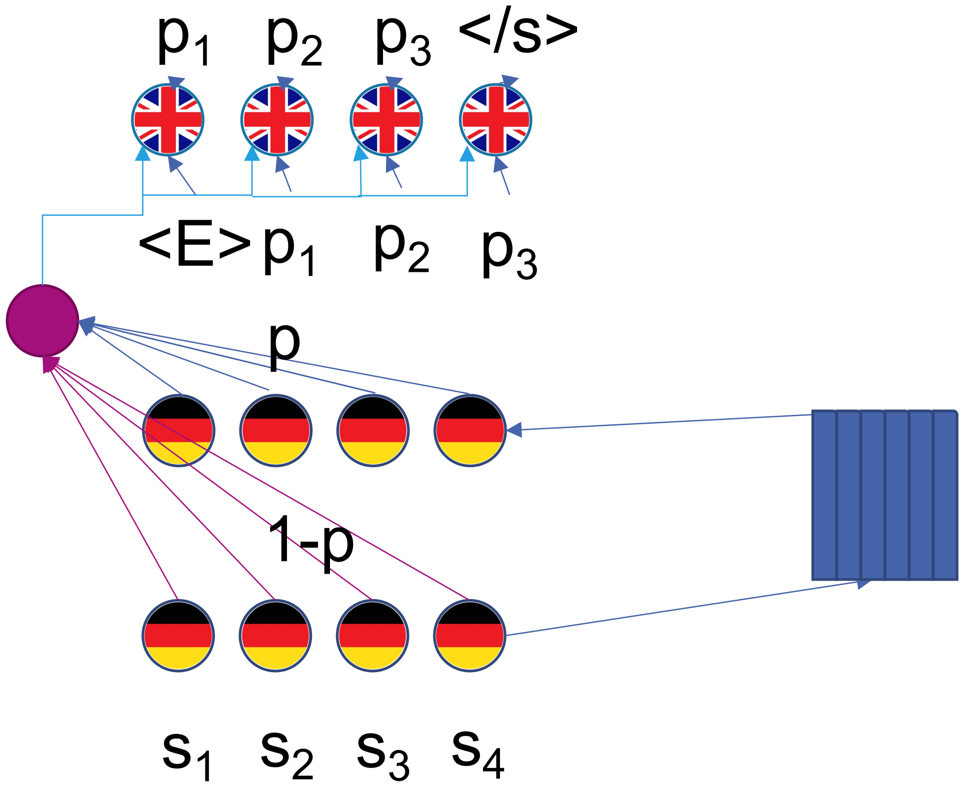
Discrete Representations

- Challenge:
  - Learning representation
  - Backpropagation
    - Straight-through estimator



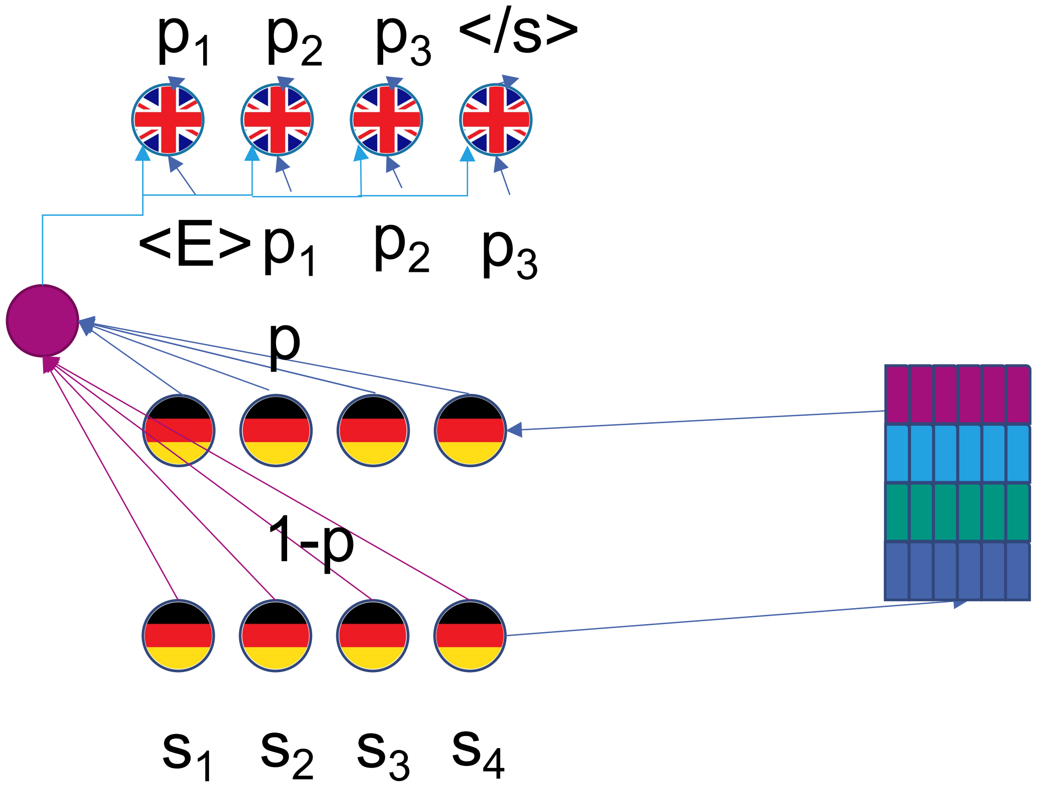
Discrete Representations

- Challenge:
  - Learning representation
  - Backpropagation
  - Less expressive
    - Information bottleneck
    - Soft discretization



Discrete Representations

- Challenge:
  - Learning representation
  - Backpropagation
  - Less expressive
  - Index collapse
    - Slicing the codebook
      - Kaiser et al. ,2018



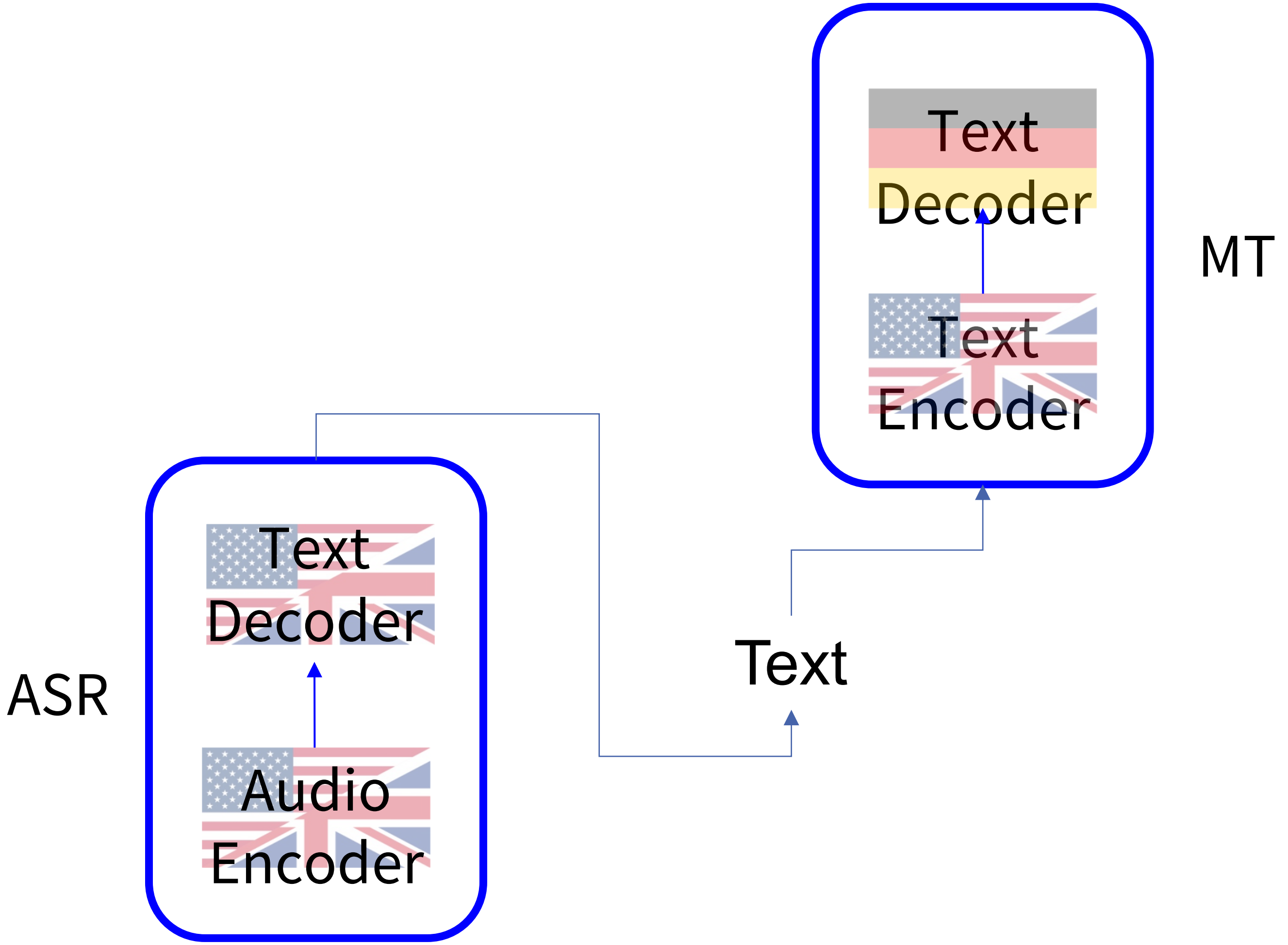
Results

- Zero-shot translation quality
  - Initialized with MM100
  - Different bridge langauges
  - BLEU Score

Dataset	Baseline	Sim	Adv	Discrete
ID-Bridge	17.7	18.4	18.4	18.3
EN-Bridge	5.1	17.3	17.2	15.2

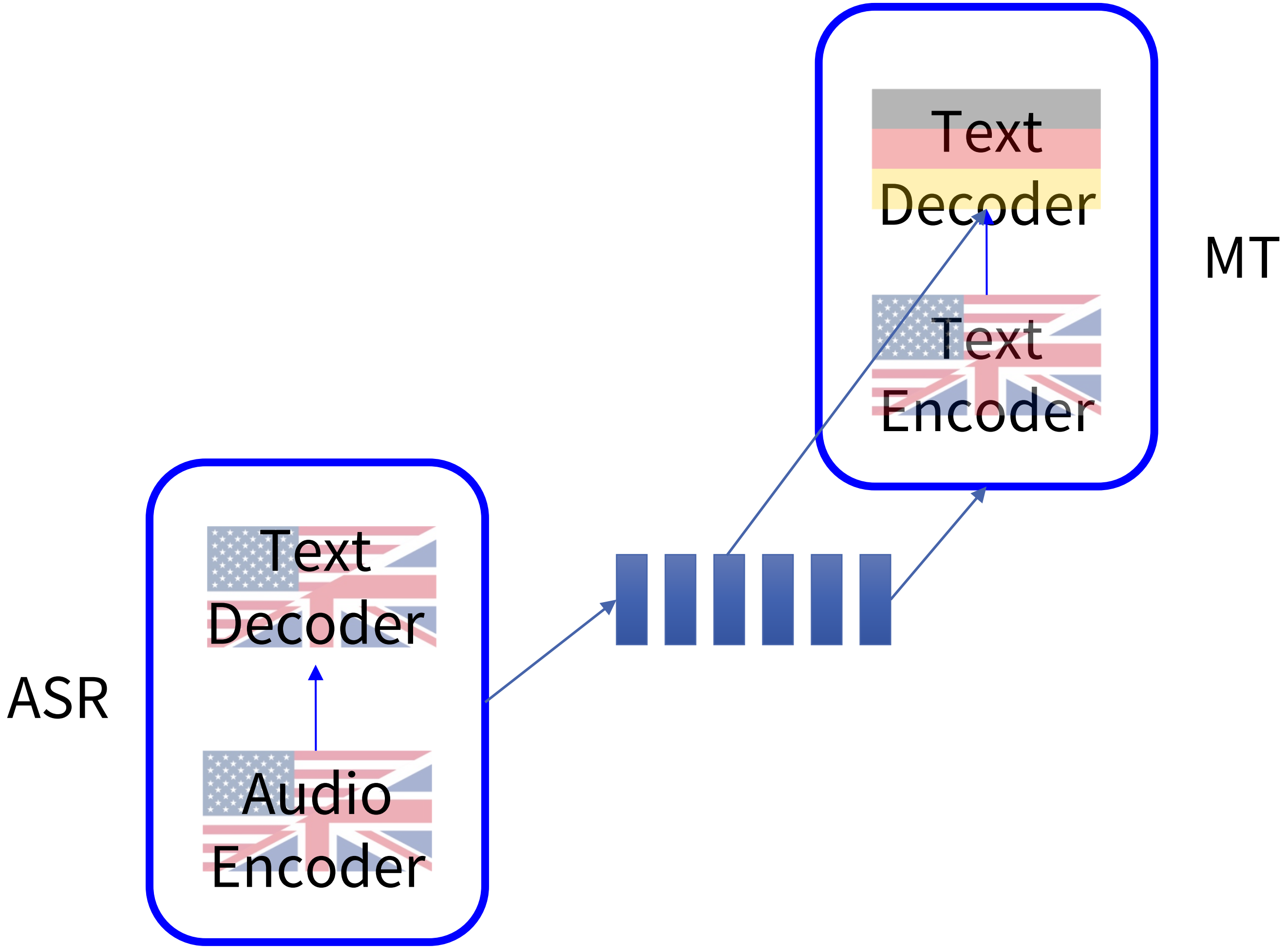
Speech Translation

- Cascaded Speech Translation
  - ASR
  - MT



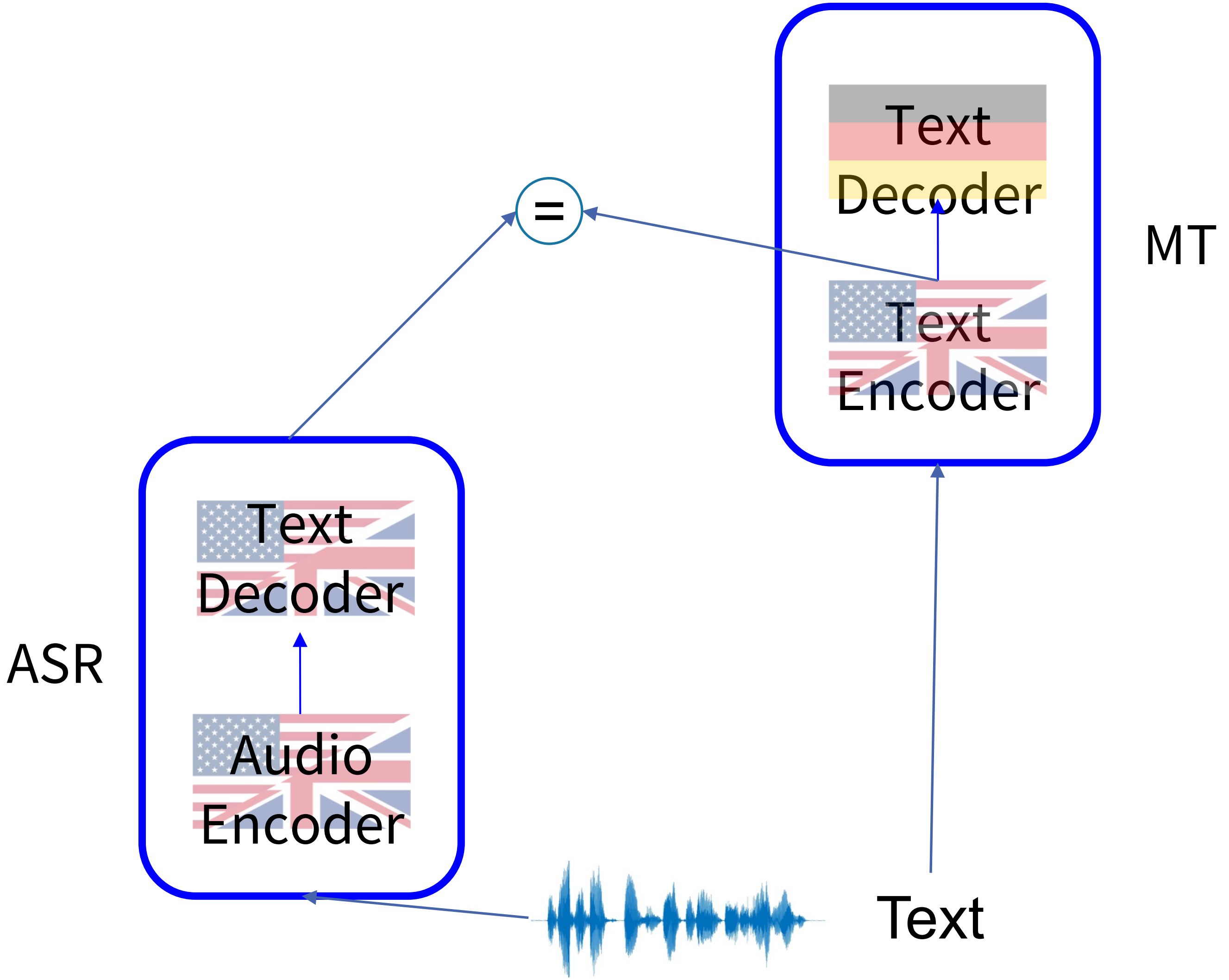
Speech Translation

- Cascaded Speech Translation
  - ASR
  - MT
- End-to-End speech translation
  - One single model
  - Mainly ASR/MT training data



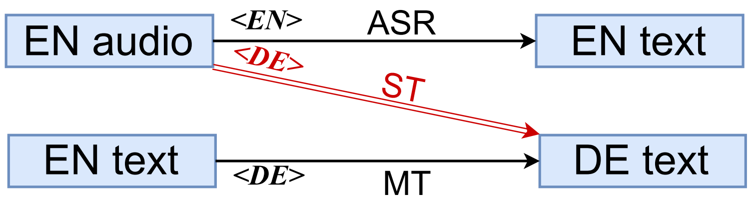
Speech Translation

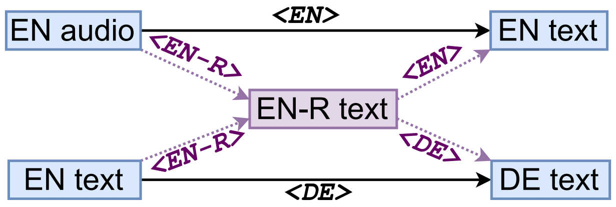
- Cascaded Speech Translation
  - ASR
  - MT
- End-to-End speech translation
  - One single model
  - Mainly ASR/MT training data
- Increase similarity



Data augmentation

- Single bridge difficult
  - Add artificial language
    - Artificial language: character-wise-reversed English (EN-R)
    - E.g. “Hello world!” → “Dlrow olleh!”

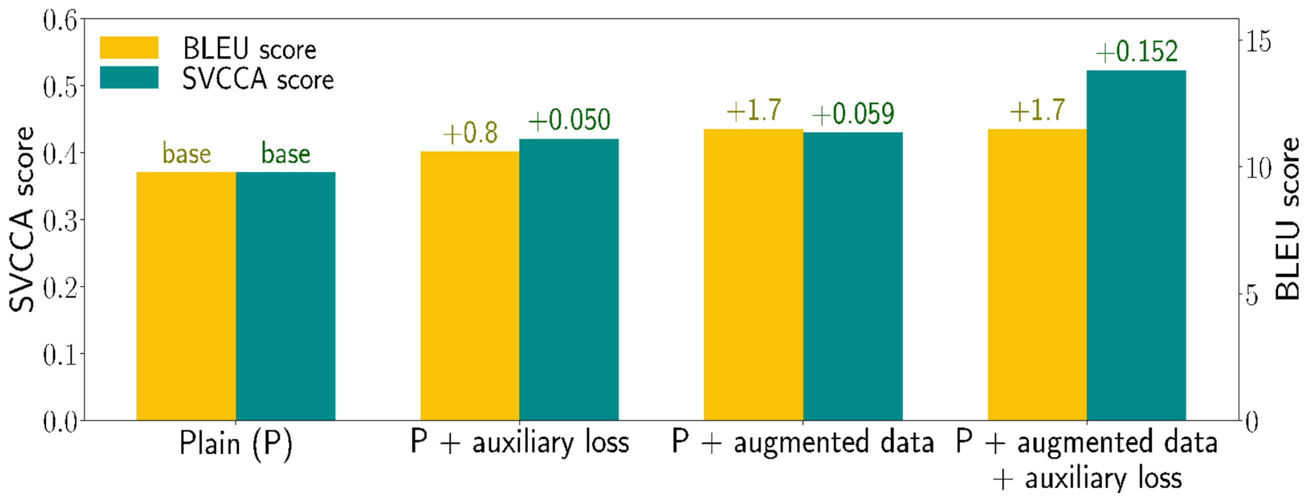




Results

	10% ST data for fine-tuning	25% ST data for fine-tuning
Plain proposed model	9.8	12.4
Plain proposed model + similarity loss	10.6 **(+0.8)**	13.2 **(+0.8)**
Plain proposed model + augmented data	11.5 **(+1.7)**	13.5 **(+1.1)**
Plain proposed model + augmented data + similarity loss	11.5 **(+1.7)**	13.7 **(+1.3)**

Results



BLEU score
SVCCA score
SVCCA score
BLEU score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0
5
10
15
base
base
+0.8
+0.050
+1.7
+0.059
+1.7
+0.152
Plain (P)
P + auxiliary loss
P + augmented data
P + augmented data
+ auxiliary loss

Results Pre-trained Models

Experiment	Without	10%	15%	20%	All
Only original loss	-	0.32	1.98	11.8	20.9
After similarity loss	0	0.98	10.7	17.8	21.6

Conclusion

- Encoder-Decoder Models assume End-to-End data
  - Often not available
- Compatibility of representation essential
- Different techniques to achieve
  - Similarity losses
  - Adversarial losses
  - Architectural changes
  - Discrete representation

References

- Li, Z., & Niehues, J. (2022). Efficient Speech Translation with Pre-trained Models. *Conference on Neural Information Processing Systems (NeurIPS) 2022*.
- Liu, D., Niehues, J., Cross, J., Guzmán, F., & Li, X. (2021). Improving Zero-Shot Translation by Disentangling Positional Information. 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), 1259–1273. https://doi.org/10.18653/v1/2021.acl-long.101
- Liu, D., Niehues, J. (2022). Learning an Artificial Language for Knowledge-Sharing in Multilingual Translation. Proceedings of the 7th Conference on Speech Translation.
- Dinh, T. A., Liu, D., & Niehues, J. (2022). Tackling Data Scarcity in Speech Translation Using Zero-Shot Multilingual Machine Translation Techniques. ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6222–6226. https://doi.org/10.1109/ICASSP43922.2022.9746815

Thanks

https://ai4lt.anthropomatik.kit.edu