Kyoto University MT System Description for IWSLT 2017

Raj Dabre[1], Fabien Cromieres[2], Sadao Kurohashi[1]

[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2]Japan Science and Technology Agency, Saitama, Japan

QR code to access Slides

Flow of This Talk

- Overview
  - Multilingual Task
  - AIAYN
- Our approaches
  - Using NMT
  - Using SMT (for internal evaluation)
- Experimental Settings
- Results and Observations
- Conclusion

Multilingual Task

- 5 languages
  - German, Dutch, Romanian, Italian and English
  - 3 Germanic and 2 Romance
- Objective: One multilingual model for all 5 languages (20 directions)
- Non zero-shot setting
  - Use all data (20 parallel corpora)
- Zero-shot setting
  - All data except for German-Dutch, Dutch-German, Romanian-Italian and Italian-Romanian (16 parallel corpora)

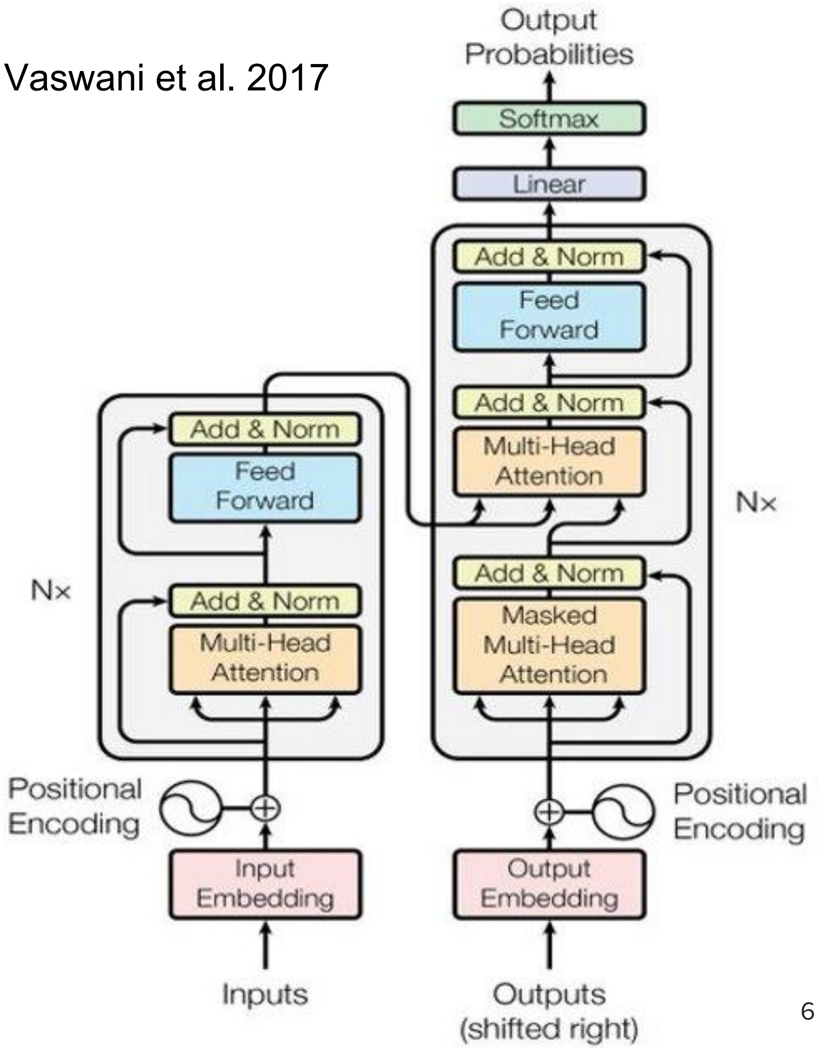
Preferred Paradigm: Non-Recurrent NMT

- Why NMT?
  - Easier to develop end-to-end multilingual models with parameter sharing (Johnson et al., 2016)
  - NMT as a black box is good enough
- Why Non-Recurrent?
  - Faster to train (multilingual model training takes time as it is)
  - Known to perform better than recurrent models (Vaswani et al., 2017)

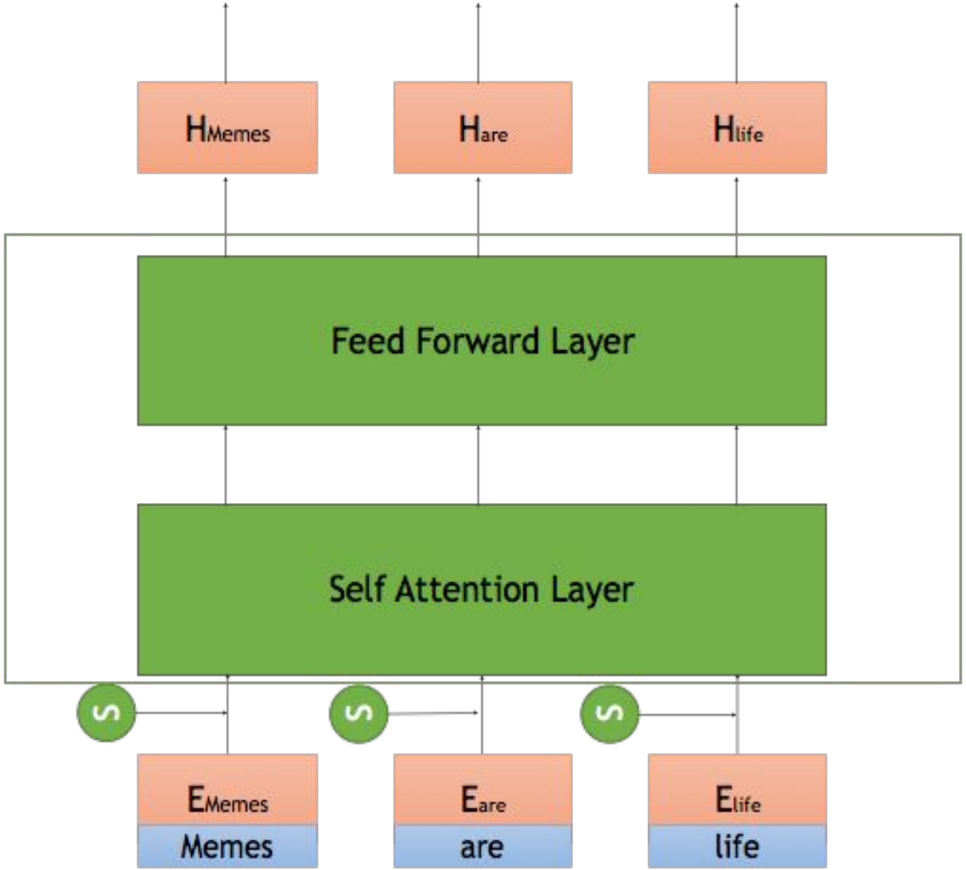
Attention is All You Need (AIAYN)

Taken from Vaswani et al. 2017

- Faster training
  - Feed Forward Layers
  - Positional Encoding
  - Residual connections
  - Batch Normalization
- Better attention mechanism
  - Multi-head
  - Self and cross
- Adam with decay

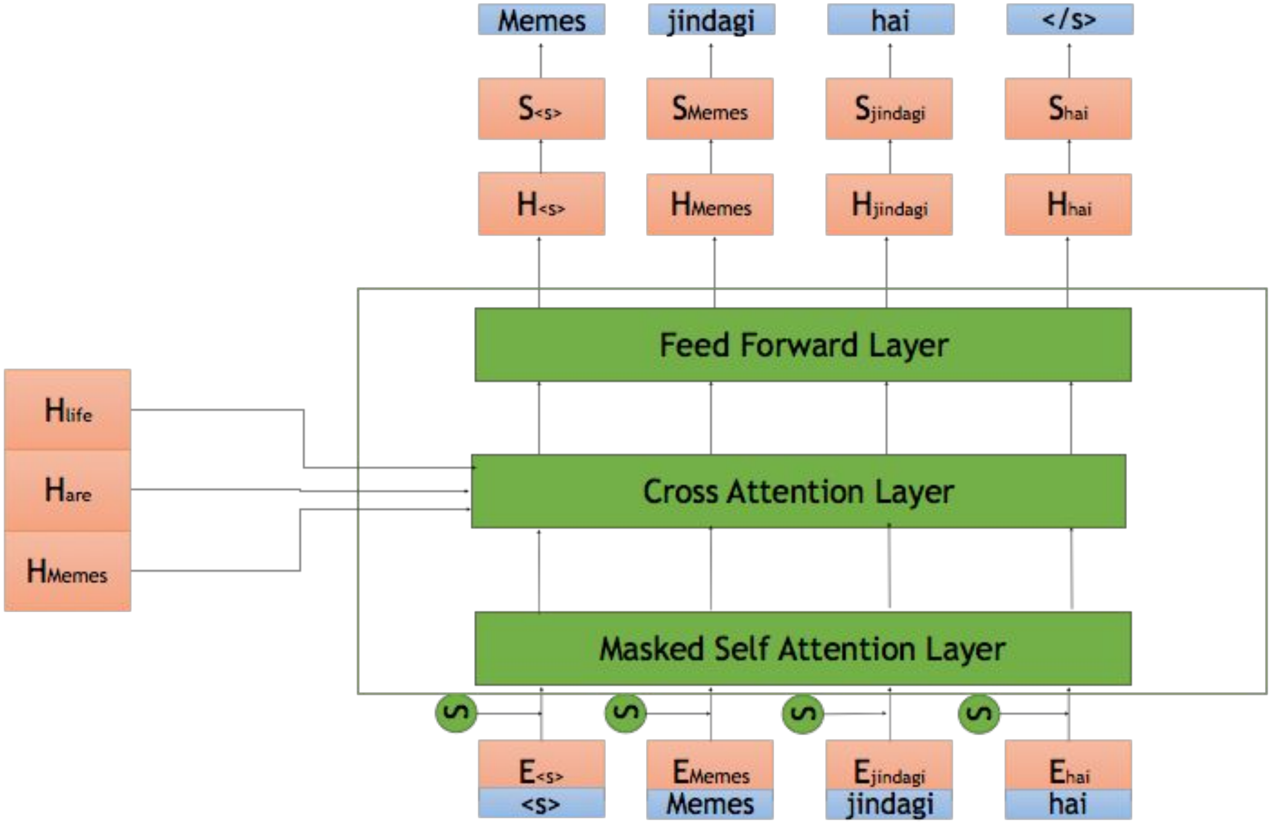


Encoder



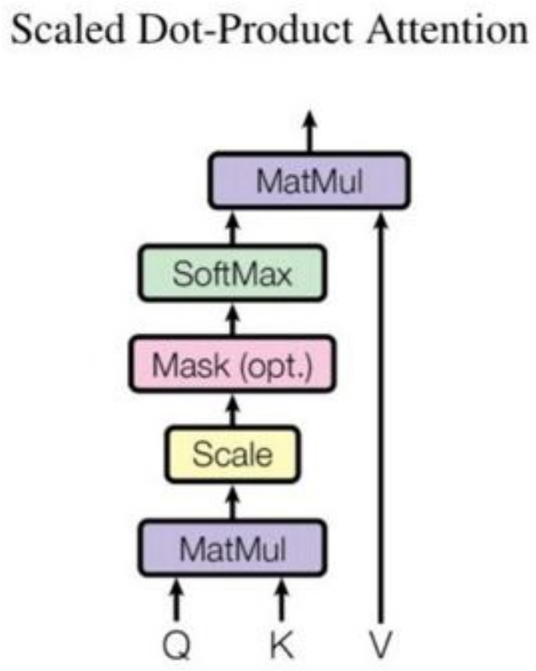
N such layers
Residual connections
Normalization

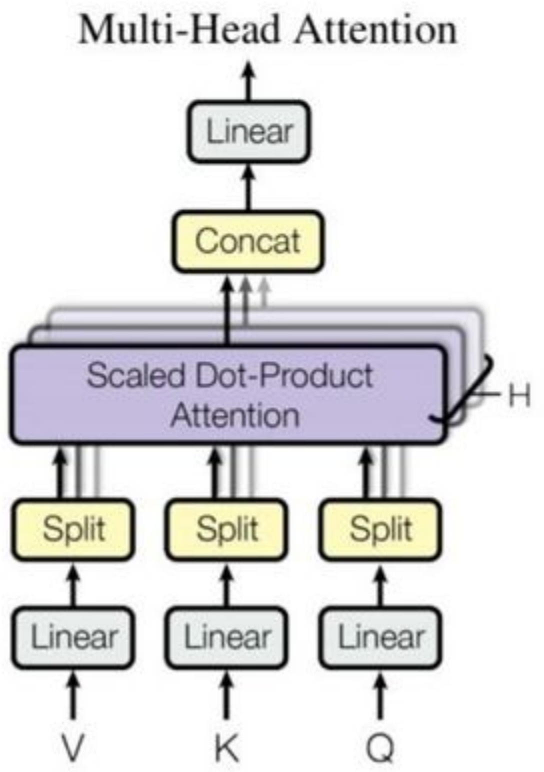
Decoder



N such layers
Residual connections
Normalization

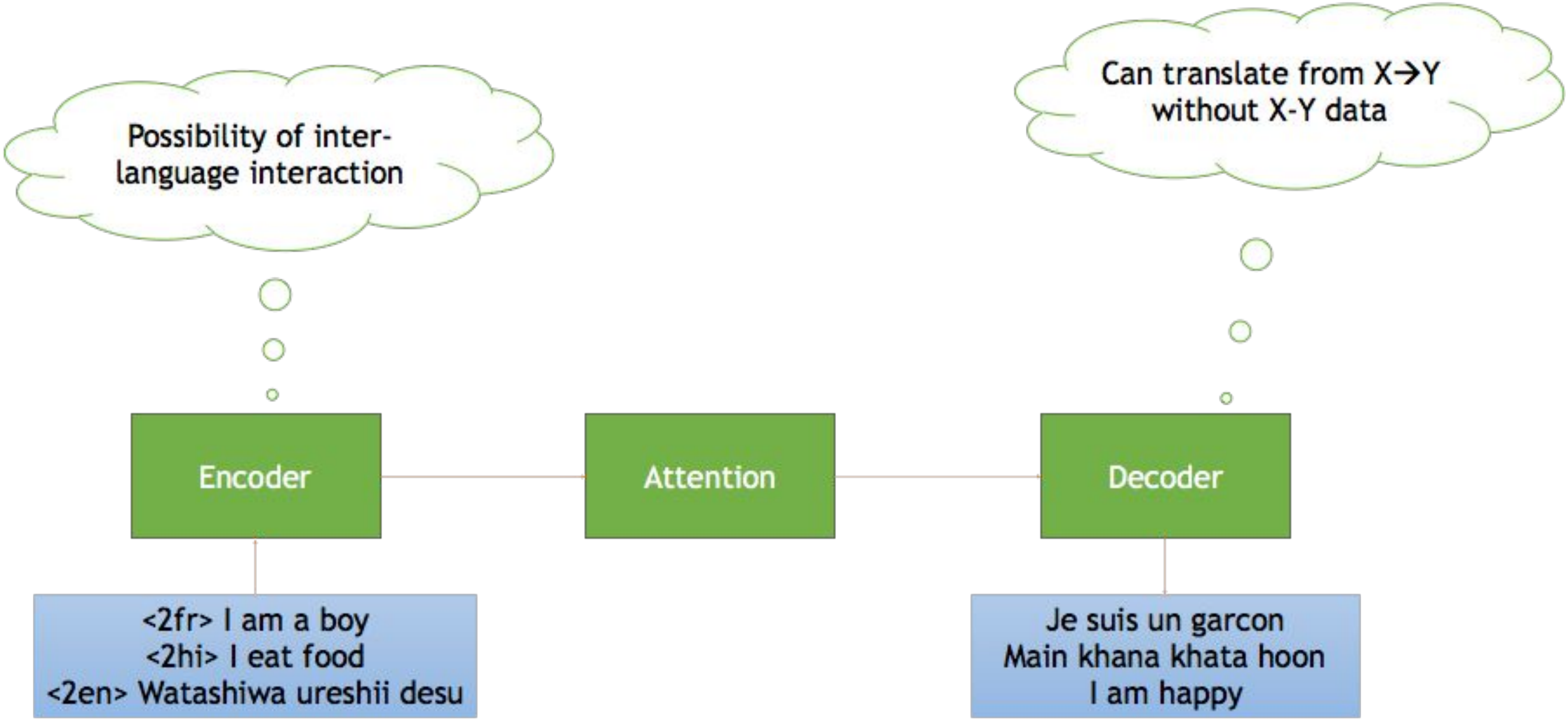
Attention





Taken from Vaswani et al. 2017

Our Approach: MLNMT Using Artificial Tokens



Multilingual Phrase based SMT

- Hacky Approach
- Only works for non-zero shot conditions
- **Technique:** For each language pair append "#tgt" to each source token
- **Example:**
  - Original: "I am a boy" --> "Watashi wa otokonoko desu"
  - Modified: "I#ja am#ja a#ja boy#ja" --> "Watashi wa otokonoko desu"
- **Outcome:** Single phrase table with multiple language directions
- **Working:** Token "#tgt" helps match phrase pairs for exactly one language pair

Experimental Settings

- Corpora
  - 20 way corpora provided by organizers (~200K sentences per direction)
  - dev2010 and tst2010 for internal evaluation
  - tst2017 for official evaluation
- Generic Preprocessing
  - XML to Moses format
  - Tokenization (using Moses tokenizer)
  - Truecasing (using Moses truecaser)
- Specific Preprocessing For NMT:
  - Prepending the “<2xx>” token to source sentences for all corpora
- Specific Preprocessing For PBSMT:
  - Appending “#xx” token to all source word tokens for all corpora
  - Byte Pair Encoding
    - Not needed for NMT: AIAYN has in built sub-word encoder

PBSMT settings

- Moses toolkit for training, tuning and testing
- Sub-word vocabulary size: 32000
- Language model: 7-gram KenLM
- Default settings for alignment and phrase extraction, tuning and testing.

NMT settings

- Google’s implementation of AIAYN
  - https://github.com/tensorflow/tensor2tensor
- Sub-word vocabulary size of 32000 (managed by EMS)
- Embedding and output layer sizes: 512
- Feed forward hidden layer size: 2018
- Adam optimizer with weight decay (Noam LR Decay)
  - 16000 of learning rate warmup before decay
- Beam search decoding:
  - Beam width of size 4
  - Alpha of 0.6 (for decoded sequence length penalty)
- Iterations: 400000 (~10 epochs)
- Data parallelism: 5 GPUs (3-4 days for convergence)

Internal Evaluation (tst2010)

- NMT is inherently superior to PBSMT
- But needs 3-4 times longer training time
- PBSMT does not really allow for languages to interact
    - No parameters are shared in reality
    - Phrase table sharing is more of a hack

Upper score is SMT
Lower score is NMT

L1/L2	de	en	it	nl	ro
de	-	29.63 34.98	17.57 21.37	23.51 23.69	14.49 18.96
en	21.70 27.81	-	24.04 29.07	27.25 30.91	21.38 26.65
it	15.88 21.37	28.89 34.58	-	18.48 21.83	19.46 20.72
nl	21.57 24.45	34.79 38.86	18.84 23.02	-	15.99 20.68
ro	15.96 21.81	31.10 37.10	22.65 24.07	18.57 23.01	-

Official Evaluation: tst2017

- **Surprise:** Zero-shot results are almost as good as non-zero shot results
- **Analysis:** Extracted 5-lingual corpora from the 20 parallel corpora
- **Observation:** 150k sentences are 5 lingual
    - **60% of corpus**
- **Conclusion:** Missing parallel sentences between Italian and Romanian and Dutch and German are remedied by indirect translations from other languages
- **Truly zero-shot?**

Non Zero Shot

L1/L2	de	en	it	nl	ro
de	-	26.45	17.54	19.64	16.27
en	23.25	-	30.79	28.80	24.66
it	19.10	34.73	-	22.32	20.60
nl	20.27	30.49	19.86	-	17.65
ro	17.94	29.58	21.89	20.24	-

Zero Shot

L1/L2	de	en	it	nl	ro
de	-	27.08	17.67	**20.31**	16.08
en	23.63	-	30.99	30.18	24.49
it	19.20	35.28	-	22.76	**20.37**
nl	**19.68**	30.63	20.74	-	17.74
ro	18.40	30.23	**21.85**	20.47	-

How does MLNMT stack against bilingual models?

- **Dutch-German**
  - Bilingual: 19.5
  - Non zero shot: **20.27**
  - Zero shot: 19.68
- **Romanian-Italian**
  - Bilingual: **23.14**
  - Non zero shot: 21.89
  - Zero shot: 21.85
- More or less comparable performance
- Bilingual models required a few hours of training on 5 GPUs

Conclusions and Future Work

- Set foundations for low resource multilingual NMT baselines
- AIAYN is fast and effective
  - Better than PBSMT setting we tried
- Zero-shot performance is almost as good or better then non zero-shot performance
  - Suspicion: Setting is not truly zero shot
- Future work
  - Train more robust models (dropout, annealing, checkpoint averaging)
  - Try out stricter zero-shot conditions
  - Better training methods for related languages (European)
  - Modifications for AIAYN for multilinguality

Thank You for Listening