UNIVERSITY OF CAMBRIDGE

The University of Cambridge's Machine Translation Systems for WMT18

Felix Stahlberg, Adria de Gispert, Bill Byrne

Overview

- Comparison of the most commonly used MT architectures
  - Neural machine translation
    - Recurrent models
    - Convolutional models
    - Self-attention-based models
  - Statistical machine translation
    - Phrase-based MT
- System combination

Data selection

English-German

Data	# Sentences
ParaCrawl	36M
All other parallel data	~4.5M

English-Chinese

Data	# Sentences
UN corpus	16M
All other parallel data	~6M

Data filtering

- **General filtering**
  - Length filtering
  - Language detection
- **ParaCrawl filtering**
  - Words with more than 40 characters
  - No HTML tags
  - 4 words minimum
  - Character ratio lower than 1:3
  - Source=target after removing non-numerical characters
  - Sentences must end with punctuation marks

Data filtering

- **General filtering**
  - Length filtering
  - Language detection
- **ParaCrawl filtering**
  - Words with more than 40 characters
  - No HTML tags
  - 4 words minimum
  - Character ratio lower than 1:3
  - Source=target after removing non-numerical characters
  - Sentences must end with punctuation marks

ParaCrawl (English-German)

	# Sentences
Original	36M
After general filtering	19M
After aggressive filtering	11M

Training data sizes

English-German

Corpus	Over-sampling	#Sentences
Common Crawl	2x	4.43M
Europarl v7	2x	3.76M
News Commentary v13	2x	0.57M
Rapid 2016	2x	2.27M
ParaCrawl	1x	11.16M
Synthetic (news-2017)	1x	20.00M
Total		**42.19M**

English-Chinese

Corpus	Over-sampling	#Sentences
CWMT - CASIA2015	2x	2.08M
CWMT - CASICT2015	2x	3.95M
CWMT - Datum2017	2x	1.93M
CWMT - NEU2017	2x	3.95M
News Commentary v13	2x	0.49M
UN v1.0	1x	14.25M
Synthetic (news-2017)	1x	20.00M
Total		**46.66M**

NMT models

- General:
  - 1024-dimensional embedding (shared on en-de) and output projection layers
  - 1024-dimensional hidden layers
  - Adam, label smoothing, layer normalization, residual connections, checkpoint averaging
  - Tokenization (Moses/Jieba), true-casing, BPE with 32K merge operations
- **LSTM**
  - 4 layers, bidirectional encoder, Bahdanau-style attention
- **SliceNet**
  - 4 convolutional layers
- **Transformer**
  - 16 head dot-product attention, 6 layers
  - absolute vs. relative positional embeddings

Architecture	en-de, de-en	zh-en
LSTM	114.2M	192.7M
SliceNet	27.5M	86.4M
Transformer	212.8M	291.4M
Relative Transformer	213.8M	292.5M

MBR-based system combination

$$S(\mathbf{y}|\mathbf{x}) = \sum_{t=1}^{T} \Big(\underbrace{\sum_{i=1}^{m} \lambda_i \log P(y_t|y_1^{t-1}, \mathbf{x}, \mathcal{M}_i)}_{\text{Full posterior}} + \underbrace{\sum_{j=m+1}^{n} \lambda_j \sum_{o=1}^{4} P(y_{t-o}^{t}|\mathbf{x}, \mathcal{M}_j)}_{\text{MBR-based } n\text{-gram scores}} \Big)$$

Why:
- Not stable
- Not possible

System combination results

	Full posterior					MBR-based n-gram scores				BLEU (test2017)		
	PBMT	LSTM	SliceNet	Trans.	Rel. Trans.	PBMT	LSTM	SliceNet	R2L Trans.	en-de	de-en	zh-en
1	✓									20.0	28.2	15.8
2		✓								28.5	35.3	23.6
3			✓							28.3	34.3	23.4
4				✓						30.5	37.9	25.6
5					✓					31.1	38.1	25.8
6				✓	✓					31.3	38.2	26.4
7		✓	✓	✓	✓					31.3	38.2	26.4
8				✓	✓		✓	✓		31.4	38.2	26.6
9				✓	✓		✓	✓	✓	31.4	38.3	26.8
10				✓	✓	✓	✓	✓	✓	31.7	38.7	27.1

2nd best | 2nd best | 7th best

Progress in MT

	English-German			German-English			Chinese-English		
	Winning system at competition	This work	Delta	Winning system at competition	This work	Delta	Winning system at competition	This work	Delta
WMT14:	20.6	31.6	11.0	29.0	36.8	7.8			
WMT15:	24.9	32.6	7.3	33.9	36.5	2.6			
WMT16:	34.2	38.5	4.3	40.2	45.1	4.9			
WMT17:	28.3	31.4	3.1	35.1	38.7	3.6	26.4	27.1	0.7
WMT18:	48.3	46.6	-1.7	48.4	48.0	-0.4	29.3	27.7	-1.6

Thanks

Training setups

Architecture	#Effective GPUs	Batch size	#SGD updates	#Training tokens
LSTM	8	4,096	45K	1,475M
SliceNet	4	2,048	800K	6,554M
R2L Transformer	16	2,048	200K	6,554M
Transformer	16	2,048	250K	8,192M
Relative Transformer	16	2,048	250K	8,192M

Single architecture results

Architecture	#Systems	English-German				German-English				Chinese-English	
		test14	test15	test16	test17	test14	test15	test16	test17	dev17	test17
PBMT	1	19.6	20.9	25.6	20.0	22.5	27.2	32.6	28.2	14.2	15.8
LSTM	1	27.1	28.8	34.6	28.0	33.8	33.3	40.7	34.8	21.8	22.7
LSTM	2	28.2	29.6	35.5	28.5	34.6	34.0	41.4	35.3	22.7	23.6
SliceNet	1	26.8	28.9	33.6	27.6	32.6	32.3	39.8	33.7	21.4	22.5
SliceNet	2	27.2	29.6	34.6	28.3	33.2	32.9	40.8	34.3	21.8	23.4
R2L Trans.	1	30.3	31.5	36.3	30.2	36.5	35.5	43.5	37.2	24.5	24.9
Transformer	1	30.7	31.9	36.6	30.5	36.7	36.2	43.7	37.9	24.9	25.6
Transformer	2	31.1	31.8	37.2	31.0	36.9	36.4	44.0	38.1	26.2	26.2
Rel. Trans.	1	31.2	31.9	37.0	31.1	37.0	36.3	44.1	38.1	24.9	25.8
Rel. Trans.	2	31.4	32.3	37.7	31.2	37.2	36.5	44.1	38.4	25.1	26.4