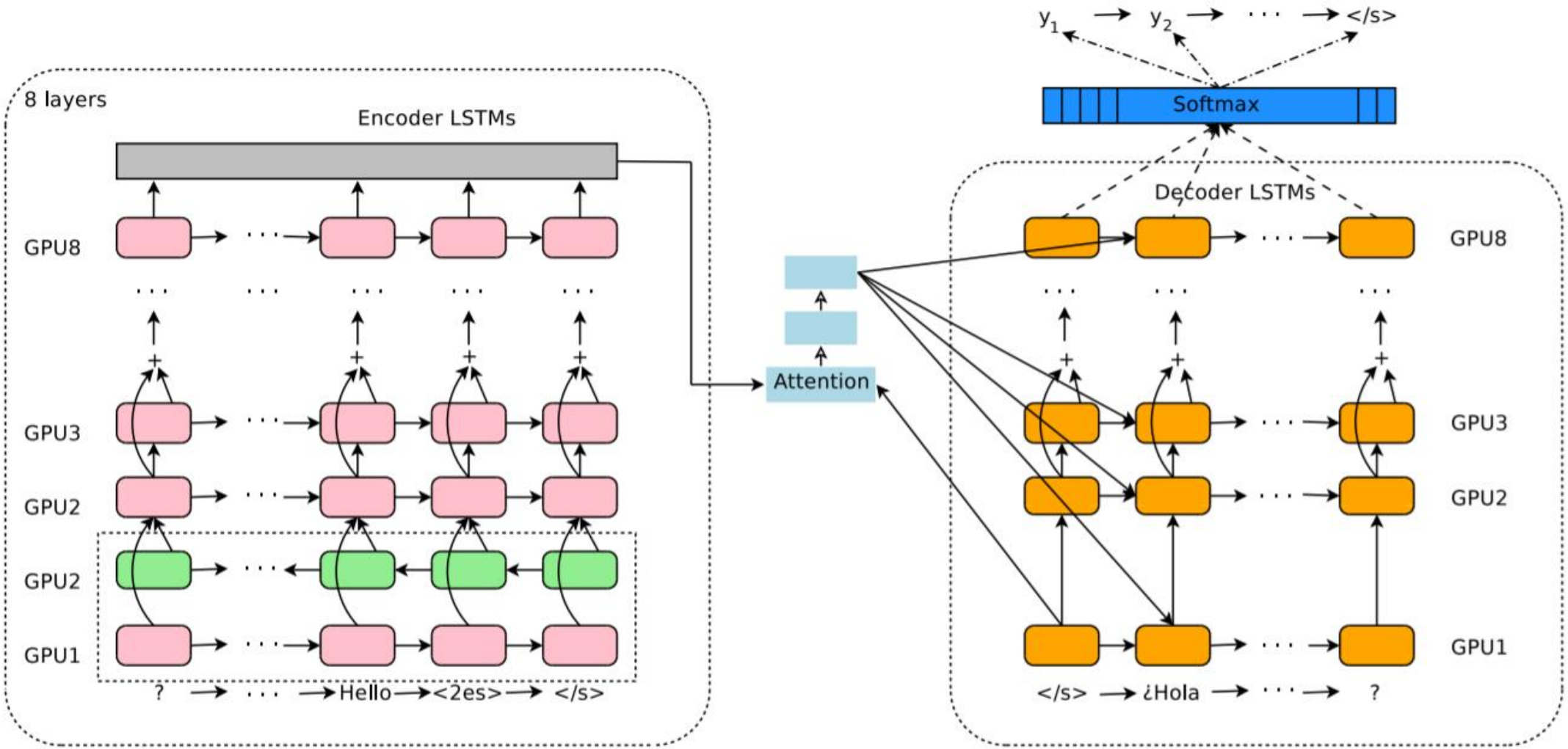
Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Mechanism

1. GNMT와 동일한 Model Architecture 및 Training Protocol을 가짐
2. 어떤 Target Language로 번역할지를 Model에게 알려 주기 위해 Artificial Token을 input의 처음에 추가
3. input의 순서를 뒤집어서 Model에게 줌
4. 하나의 Dataset을 여러 Language Pair의 Instance들로 구성하여 Catastrophic Forgetting을 방지
5. Language Pair 별 Data의 양의 차이는 Oversampling/ Undersampling을 통해 조절

1. Model Architecture



2. Contribution

1. Low-Resource Language Improvements

2. Simplicity

3. Zero Shot Translation

Low Resource Language Improvement

3. Low Resource Language Improvements

1. Resource는 사용할 수 있는 컴퓨팅 자원(CPU, GPU, Memory 등)를 뜻함

2. Low Resource Language Improvement라는 것은 컴퓨팅 자원을 적게 쓰면서도 번역 품질을 높일 수 있다는 것을 의미함.

3. Resource를 적게 사용할 수 있는 것은 개별 Language Pair 별로 NMT Model을 만드는 것보다 이러한 작업을 모두 수행할 수 있는 하나의 NMT Model을 만드는 것이 더 적은 Parameter를 사용할 수 있기 때문임.

4. 논문에서는 Improvement라는 단어를 써서 모든 경우에 대해 Multilingual GNMT Model의 번역의 품질이 기존의 Single Language Pair에 대한 번역의 품질보다 우수할 것이라고 생각할 수 있지만 실험 결과를 보면 그렇지는 않음

3. Low Resource Language Improvements - Experiments

Many to One

1. 입력으로 들어오는 Source Sentence는 여러 Language로 표현 된 것일 수 있지만 번역되어야 하는 즉 출력되어야 하는 Target Sentence는 하나의 Language로만 표현되어야 함

2. Multi Source Translation과 달리 한 번에 하나의 Source Sentence를 입력으로 받음

3. Target Language가 하나이기 때문에 Artificial Token을 추가할 필요가 없음

4. BLEU Score를 평가 Metric으로 함

5. Multi Lingual NMT Model은 Single Language Pair NMT Model과 동일한 Parameter의 수를 가짐

3. Low Resource Language Improvements - Experiments

Table 1: Many to One: BLEU scores on various data sets for single language pair and multilingual models.

Model	Single	Multi	Diff
WMT German→English (oversampling)	30.43	30.59	+0.16
WMT French→English (oversampling)	35.50	35.73	+0.23
WMT German→English (no oversampling)	30.43	30.54	+0.11
WMT French→English (no oversampling)	35.50	36.77	+1.27
Prod Japanese→English	23.41	23.87	+0.46
Prod Korean→English	25.42	25.47	+0.05
Prod Spanish→English	38.00	38.73	+0.73
Prod Portuguese→English	44.40	45.19	+0.79

1. 모든 실험에서 Multi Lingual Model이 Single Language Pair Model에 비해 높은 BLEU Score를 가짐 -> Decoder가 더 많은 English Data를 경험하기 때문

2. Dataset이 작은 경우(German) Oversampling이 도움이 되지만 Dataset이 큰 경우에는 성능을 하락시킴

3. 다른 언어에 비해 한국어, 일본어의 BLEU Score가 낮은 것은 언어 사이의 차이 때문

3. Low Resource Language Improvements - Experiments

One to Many

1. 입력으로 들어오는 Source Sentence는 하나의 Language로만 표현 되어야 하지만 번역되어야 하는 즉 출력되어야 하는 Target Sentence는 여러 Language로 표현될 수 있음

2. Target Language가 다수이기 때문에 Source Sentence에 Artificial Token을 추가해야 함

3. BLEU Score를 평가 Metric으로 함

4. Multi Lingual NMT Model은 Single Language Pair NMT Model과 동일한 Parameter의 수를 가짐

3. Low Resource Language Improvements - Experiments

Table 2: One to Many: BLEU scores on various data sets for single language pair and multilingual models.

Model	Single	Multi	Diff
WMT English→German (oversampling)	24.67	24.97	+0.30
WMT English→French (oversampling)	38.95	36.84	-2.11
WMT English→German (no oversampling)	24.67	22.61	-2.06
WMT English→French (no oversampling)	38.95	38.16	-0.79
Prod English→Japanese	23.66	23.73	+0.07
Prod English→Korean	19.75	19.58	-0.17
Prod English→Spanish	34.50	35.40	+0.90
Prod English→Portuguese	38.40	38.63	+0.23

1. 일부 실험에서는 Multi Lingual GNMT가 우수한 성능을 보이지만 Many To One과 같이 모든 실험에서 그런 것은 아님 -> 하나의 Vocab을 공유하는 상황에서 처리해야 하는 Language가 늘어난 것에 대해 Encoder는 크게 어렵다고 느끼지 않았지만 Decoder는 어려움을 느끼는 것으로 보임

2. Dataset이 작은 경우(German) Oversampling이 도움이 되지만 Dataset이 큰 경우에는 성능을 하락시킴

3. 다른 언어에 비해 한국어, 일본어의 BLEU Score가 낮은 것은 언어 사이의 차이 때문

3. Low Resource Language Improvements - Experiments

Many to Many

1. 입력으로 들어오는 Source Sentence는 하나의 Language로만 표현 되어야 하지만 번역되어야 하는 즉 출력되어야 하는 Target Sentence는 여러 Language로 표현될 수 있음

2. Target Language가 다수이기 때문에 Source Sentence에 Artificial Token을 추가해야 함

3. BLEU Score를 평가 Metric으로 함

4. Multi Lingual NMT Model은 Single Language Pair NMT Model과 동일한 Parameter의 수를 가짐

3. Low Resource Language Improvements - Experiments

Table 3: Many to Many: BLEU scores on various data sets for single language pair and multilingual models.

Model	Single	Multi	Diff
WMT English→German (oversampling)	24.67	24.49	-0.18
WMT English→French (oversampling)	38.95	36.23	-2.72
WMT German→English (oversampling)	30.43	29.84	-0.59
WMT French→English (oversampling)	35.50	34.89	-0.61
WMT English→German (no oversampling)	24.67	21.92	-2.75
WMT English→French (no oversampling)	38.95	37.45	-1.50
WMT German→English (no oversampling)	30.43	29.22	-1.21
WMT French→English (no oversampling)	35.50	35.93	+0.43
Prod English→Japanese	23.66	23.12	-0.54
Prod English→Korean	19.75	19.73	-0.02
Prod Japanese→English	23.41	22.86	-0.55
Prod Korean→English	25.42	24.76	-0.66
Prod English→Spanish	34.50	34.69	+0.19
Prod English→Portuguese	38.40	37.25	-1.15
Prod Spanish→English	38.00	37.65	-0.35
Prod Portuguese→English	44.40	44.02	-0.38

3. Low Resource Language Improvements - Experiments

Table 4: Large-scale experiments: BLEU scores for single language pair and multilingual models.

Model	Single	Multi	Multi	Multi	Multi
#nodes	1024	1024	1280	1536	1792
#params	3B	255M	367M	499M	650M
Prod English→Japanese	23.66	21.10	21.17	21.72	21.70
Prod English→Korean	19.75	18.41	18.36	18.30	18.28
Prod Japanese→English	23.41	21.62	22.03	22.51	23.18
Prod Korean→English	25.42	22.87	23.46	24.00	24.67
Prod English→Spanish	34.50	34.25	34.40	34.77	34.70
Prod English→Portuguese	38.40	37.35	37.42	37.80	37.92
Prod Spanish→English	38.00	36.04	36.50	37.26	37.45
Prod Portuguese→English	44.40	42.53	42.82	43.64	43.87
Prod English→German	26.43	23.15	23.77	23.63	24.01
Prod English→French	35.37	34.00	34.19	34.91	34.81
Prod German→English	31.77	31.17	31.65	32.24	32.32
Prod French→English	36.47	34.40	34.56	35.35	35.52
ave diff	-	-1.72	-1.43	-0.95	-0.76
vs single	-	-5.6%	-4.7%	-3.1%	-2.5%

Simplicity

4. Simplicity

1. Multi Lingual NMT를 위해 Model의 Architecture에 어떠한 변화도 주지 않음

2. Model의 Dataset만을 조정하여 Multi Lingual NMT Model을 학습시킬 수 있음
   -여러 Parallel Corpus를 합쳐 하나의 Dataset을 구성
   -Source Sentence에 첫 부분에 Artificial Token을 추가
   -Language 별 데이터 양을 근거로 Oversampling과 Downsampling을 수행

3. Machine Translation Service의 배포 난이도를 낮춤
   -Model 구축을 위해 필요한 Parameter의 수를 줄여 Resource에 대한 제약을 줄임
   -번역 요청으로 들어온 여러 언어로 이루어진 Source Sentence를 하나의 Mini Batch로 묶어 처리할 수 있게 되어 Parallelism을 높일 수 있음

4. Catastrophic Forgetting 문제를 해결하기 위해 기존에는 Training 과정에서 복잡한 Scheduling을 해야 했으나 더 이상 이러한 Scheduling을 하지 않아도 됨

Zero Shot Translation

5. Zero Shot Translation

1. Multi Lingual NMT Model에 대하여 학습 과정에서 주어지지 않은 Language Pair에 대해서도 일정 수준 이상으로 번역이 가능

2. NMT Model에서 Transfer Learning이 이루어졌다는 증거
   - Transfer Learning은 사전에 일반적인 지식을 학습시키고 이후 더 전문적인 Task를 수행하는 방법에 대한 학습을 수행함으로써 모델의 성능을 높이고자 하는 방법

3. Zero Shot Translation vs Zero Resource Translation
   - Zero Resource Translation은 Pretraining된 Multi Lingual NMT Model을 Fine Tuning할 때 실제 데이터가 아니라 Multi Lingual NMT Model에 의해 생성된 Pseudo Parallel Data로 Fine Tuning한 후에 Translation을 수행 (여기서 Resource는 Data를 의미)

   - Zero Shot Translation은 Multi Lingual NMT Model을 Fine Tuning하지 않고 Translation을 수행

5. Zero Shot Translation – Effect of Direct Parallel Data

1. 이용 가능한 parallel dataset을 기반으로 학습의 효과를 높이기 위한 방법에 대한 실험
   -Pretraining을 하고 Fine Tuning을 하는 방법(Incremental)

   -처음 부터 모든 Dataset을 사용하여 (From-Scratch)

2. Experiment Setup
   -Zero Shot(Baseline) : Dataset(En <-> {Belarusian, Russian, Ukranian}), Oversampling

   -From Scratch : Dataset(En <-> {Belarusian, Russian, Ukranian} + Ru <-> {Be, Uk}),
   Oversampling

   -Incremental : Zero Shot(Dataset(En <-> {Belarusian, Russian, Ukranian}))
   + Fine Tuning(Data from 'From-Scratch')

5. Zero Shot Translation – Effect of Direct Parallel Data

Table 7: BLEU scores for English↔{Belarusian, Russian, Ukrainian} models.

	Zero-Shot	From-Scratch	Incremental
English→Belarusian	16.85	17.03	16.99
English→Russian	22.21	22.03	21.92
English→Ukrainian	18.16	17.75	18.27
Belarusian→English	25.44	24.72	25.54
Russian→English	28.36	27.90	28.46
Ukrainian→English	28.60	28.51	28.58
Belarusian→Russian	56.53	82.50	78.63
Russian→Belarusian	58.75	72.06	70.01
Russian→Ukrainian	21.92	25.75	25.34
Ukrainian→Russian	16.73	30.53	29.92

1. 구분선 위의 것은 영어를 Source/Target Language로 하는 경우고 구분선 아래는 영어 이외의 언어를 Source/Target Language로 하는 경우

2. Baseline인 Zero Shot도 일정 수준 이상의 성능을 보임

3. 학습 과정에서 경험한 Task에 대해서는 Incremental Model이 From-Scratch보다 높으나 그렇지 않은 경우에는 From-Scratch가 Incremental보다 높음

5. Zero Shot Translation – Visual Analysis

1. 목적
   (1).Multi Lingual Network가 Shared Representation을 학습할 수 있는지를 확인

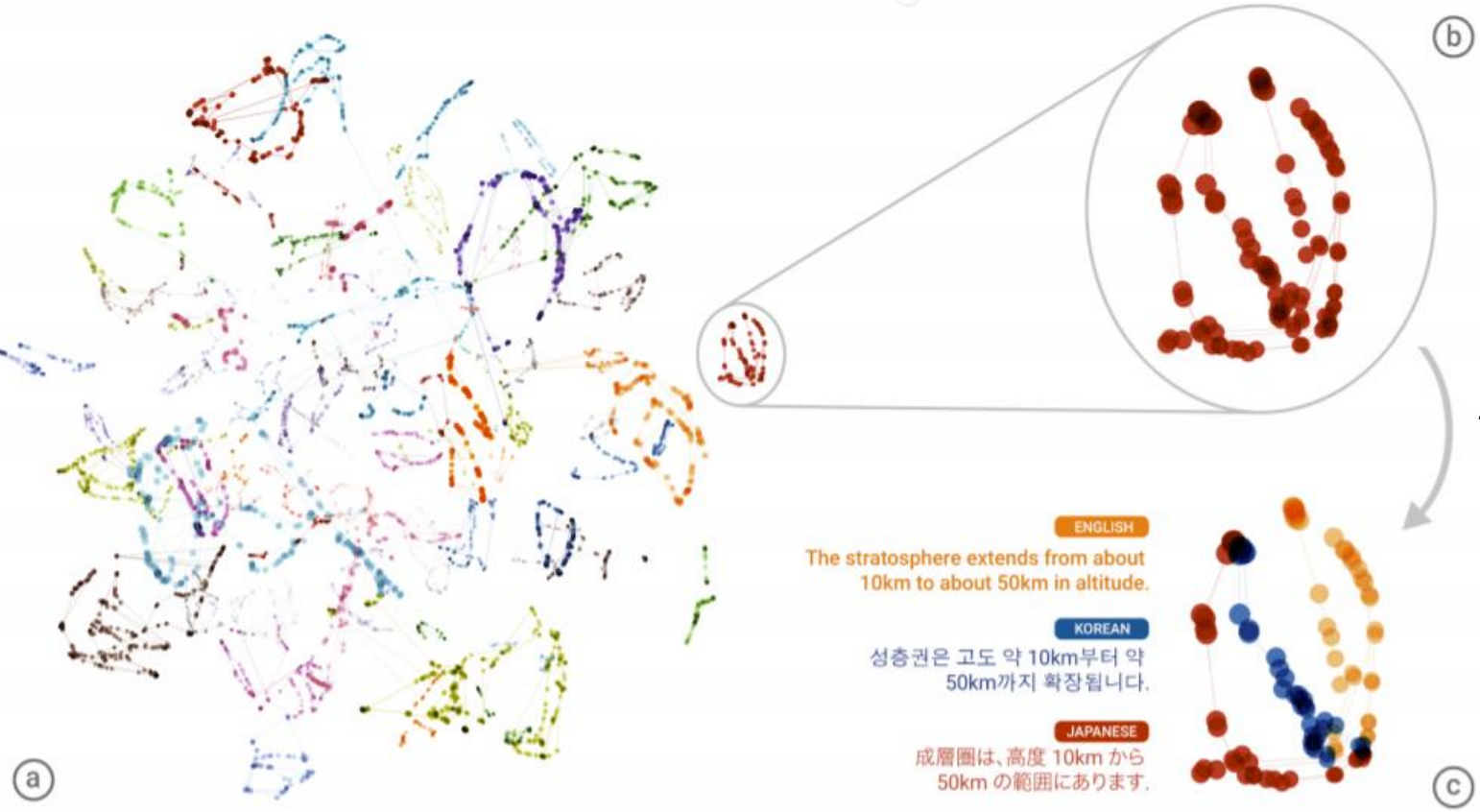
   (2).학습 과정에서 경험하지 못한 Language Pair에 대해서도 학습 과정에서 경험한 Language Pair와 동일한 방식으로 취급하는지를 확인

2. Context Vector 사이의 상관 관계를 분석
   -언어에 관계 없이 유사한 의미의 input이 들어왔을 때 유사한 context vector가 나타난다면 이는 목적 (1)을 뒷받침하는 근거가 될 수 있음

   -모델이 학습 과정에서 경험하지 못한 Language Pair와 학습 과정에서 경험한 Language Pair에 대해서 유사한 의미의 input이 들어왔을 때 유사한 context vector가 나타난다면 이는 목적 (2)를 뒷받침하는 근거가 될 수 있음

3. Context Vector 사이의 상관 관계 분석을 시각화하기 위해 벡터 사이의 거리를 유지하면서도 차원을 축소시켜주는 t-SNE 방법을 기반으로 한 TensorFlow Embedding Projector를 사용

5. Zero Shot Translation – Evident for an Interlingua



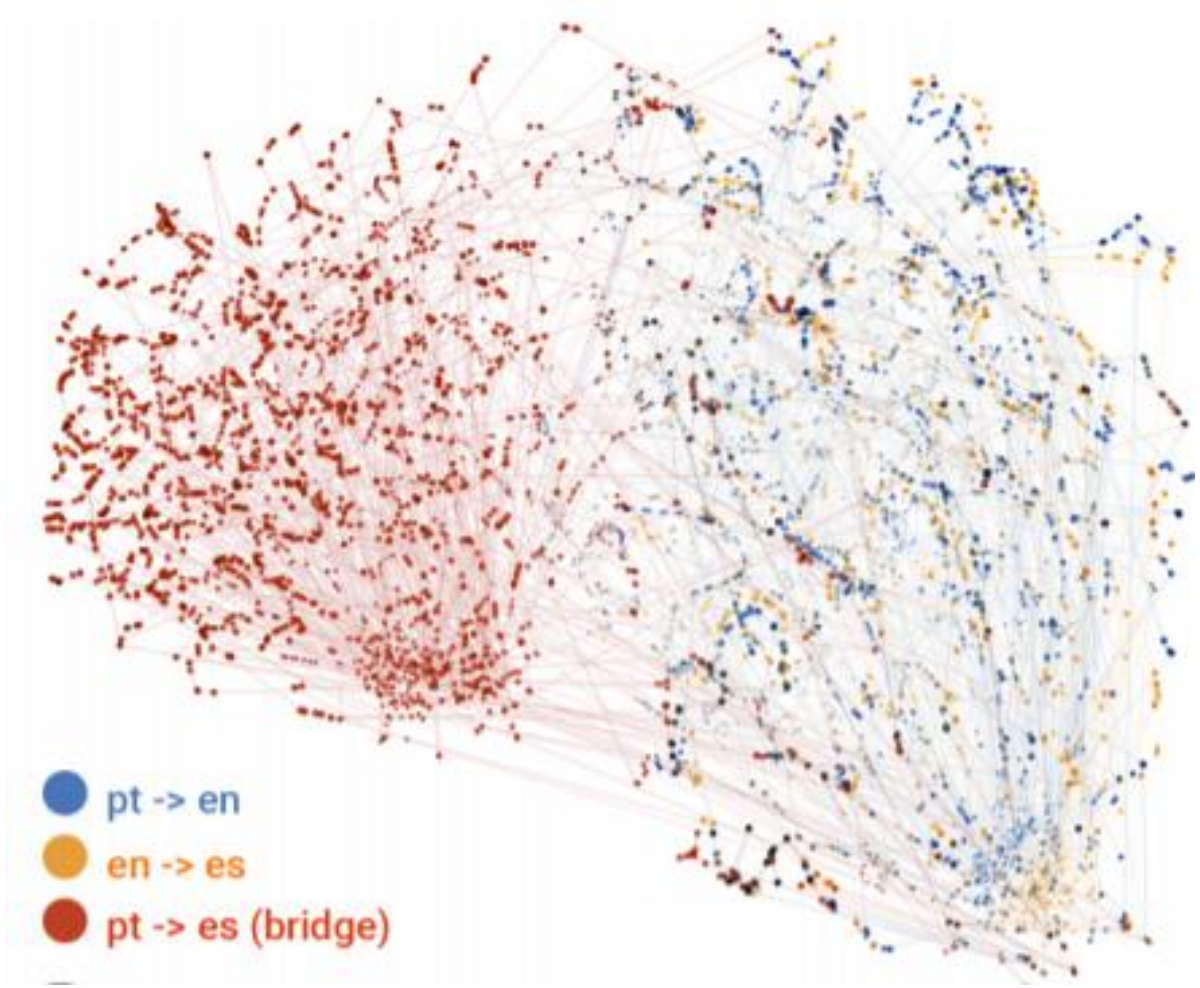
1. 각 Point는 Decoding Timestep 별 context vector를 축소한 것이며 하나의 triplet에 속하는 context vector들은 같은 색으로 칠해져 있으며 선으로 연결되어 있음

2. Context vecto,r는 학습 과정에서의 경험 여부에 관계 없이 모든 Languaga Pair에 대해 task를 수행하는 과정에서 계산된 것

4. 왼쪽 그림을 보면 같은 색의 Point들이 밀집하여 위치해 있는 걸로 보아 언어가 달라도 의미가 같으면 model이 유사한 context vector를 계산한다고 판단할 수 있음

5. Model은 En <-> Ja, En <-> Ko Task에 대하여 학습을 수행

5. Zero Shot Translation – Partially Separated Representations



1. 각 Point는 Decoding 과정에서 계산된 각 Timestep 별 context vector를 축소한 것이며 하나의 triplet에 속하는 context vector들은 같은 색으로 칠해져 있으며 선으로 연결되어 있음

2. Model은 Po -> En, En -> Es Task에 대하여 학습을 수행

3. 왼쪽 그림을 보면 학습 과정에서 경험한 Task의 context vector인 파란색과 노란색의 point들은 비슷한 위치에 분포하고 있으나 학습 과정에서 경험하지 못한 Task인 Pt ->Es Task에 대한 context vector는 이들과 완전히 분리되어 있음

5. Zero Shot Translation – Partially Separated Representations

1. Partially Separated Representation이 실제 Translation Quality에 영향을 미치는지에 대해 확인하고자 실험을 수행

2. BLEU Score와 논문에서 자체적으로 정의한 Dissimilairty Measure 사이의 상관 관계를 통해 이를 확인하고자 함. 여기서 BLEU Score는 Translation Quality를 정량화한 값이며 Dissimilarity Measure는 다른 언어로 표현된 동일한 의미의 문장 사이의 길이의 차이가 미치는 영향을 배제한 두 문장 사이의 유사도를 정량화한 값

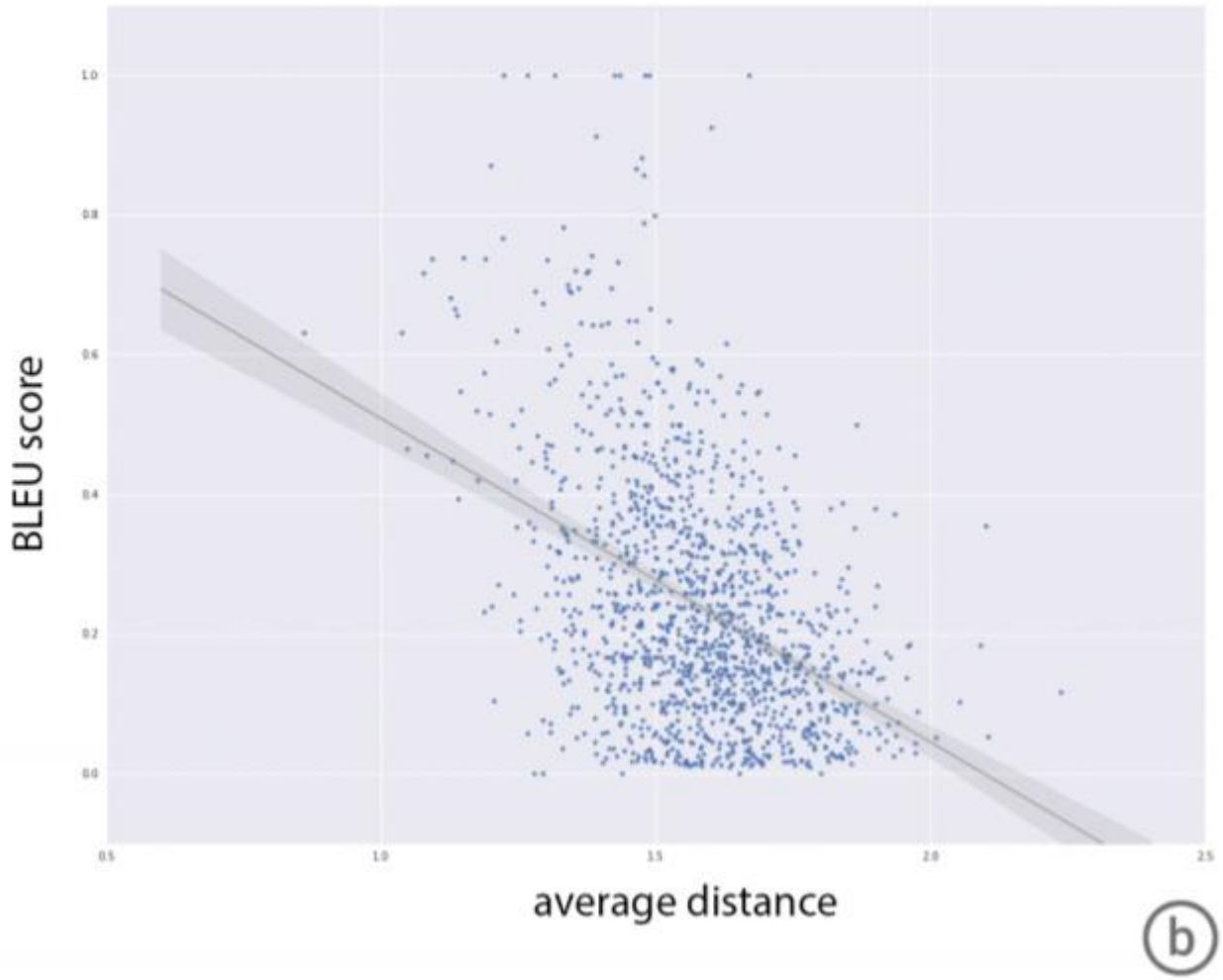
5. Zero Shot Translation – Partially Separated Representations

$$dissimilarity(\gamma_1, \gamma_2) = \frac{1}{m}\sum_{i=0}^{m-1} d\left(\gamma_1\left(\frac{i}{m-1}\right), \gamma_2\left(\frac{i}{m-1}\right)\right)$$

$$\gamma\left(\frac{i}{n-1}\right) = y_i$$

1. m = 두 문장의 길이 중 더 큰 값
2. $\gamma(x)$은 실수를 입력받아 벡터를 출력하는 함수
3. $\gamma_1(x)$와 $\gamma_2(x)$는 동일한 의미이지만 표현된 언어가 다른 문장을 나타냄
4. 주어진 두 개의 문장중 길이가 짧은 문장은 위의 수식을 계산하는 과정에서 i가 특정 값 이상이 되면 더 이상 값이 존재하지 않기 때문에 원칙적으로는 계산을 수행할 수 없는데 이를 해결하기 위해 linear interpolation을 사용

5. Zero Shot Translation – Partially Separated Representations



1. 가로 축은 Dissimilarity Measure이고 세로 축은 BLEU Score

2. 그래프의 직선을 보면 음의 상관 관계가 있음을 확인할 수 있음

3.Pearson Coefficient를 구하면 -0.42로 두 값 사이에 음의 상관 관계가 있음을 확임할 수 있음

4. 따라서 동일한 의미의 다른 언어로 표현된 문장에 대한 Representation 사이의 거리가 멀수록 Translation Quality가 떨어진다고 판단할 수 있음

5. 즉 Multi Lingual NMT Model의 Translation Quality는 Model이 얼마나 Shared Representation을 잘 학습하는지에 따라 달라진다.

6. Extras

Mixing Languages – Source Sentence

1. Code Switching : Source Language의 하나의 문장을 여러 언어를 혼용하여 표현하는 것

2. 아래의 예시에서는 Model이 크게 다르지 않게 번역하였으나 번역을 제대로 수행하지 못하는 경우도 다수 있다고 논문에서 언급

- **Japanese:** 私は東京大学の学生です。 → I am a student at Tokyo University.
- **Korean:** 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.
- **Mixed Japanese/Korean:** 私は東京大学학생입니다. → I am a student of Tokyo University.

6. Extras

Mixing Languages – Artificial Token

1. 실험 결과를 보면 일부 사례에서는 출력 값에 두 언어가 혼재하기도 하지만 대부분의 경우 w가 대략 0.5 정도가 되면 아예 다른 언어로 바뀜
2. 이를 통해 Model이 두 가지 언어를 혼용하는 것을 어려워한다는 가설을 세울 수 있음
3. Artificial Token이 문장의 첫 부분에 위치해 있기 때문에 Decoding이 진행되면 진행 될 수록 Decoder는 어떤 문장을 번역할 언어를 결정할 때 Artificial Token 보다는 이전 Time step에 의존하기 때문에 일정 Time Step이후로는 오직 하나의 언어로만 번역됨. 반대로 이전에는 Artificial Token의 영향을 받기 때문에 Artificial Token이 섞인 경우 번역 결과도 섞일 수 있음.

Japanese/Korean:	I must be getting somewhere near the centre of the earth.
$w_{ko} = 0.00$	私は地球の中心の近くにどこかに行っているに違いない。
$w_{ko} = 0.40$	私は地球の中心近くのどこかに着いているに違いない。
$w_{ko} = 0.56$	私は地球の中心の近くのどこかになっているに違いない。
$w_{ko} = 0.58$	私は지구の中心의가까이에어딘가에도착하고있어야한다。
$w_{ko} = 0.60$	나는지구의센터의가까이에어딘가에도착하고있어야한다。
$w_{ko} = 0.70$	나는지구의중심근처어딘가에도착해야합니다。
$w_{ko} = 0.90$	나는어딘가지구의중심근처에도착해야합니다。
$w_{ko} = 1.00$	나는어딘가지구의중심근처에도착해야합니다。

6. Extras

Input의 형식 관한 의문과 설명을 위한 가설

1. 왜 Input의 순서를 뒤집는 걸까?
   -Artificial Token을 뒤에 놓기 위함

2. 왜 Artificial Token을 뒤에 놓아야 할까?
   -Artificial Token을 뒤에 놓는 것은 벡터가 담을 수 있는 정보량의 한계로 인해 Artificial Token을 앞에 놓으면 Encoding Timestep이 뒤로 갈수록 어떤 언어로 번역해야 할지에 관한 정보가 소실되기 때문인 듯.

   -Attention Mechanism을 사용하기 때문에 이는 문제가 안되는 것 아닌가 싶지만 Mixing Language 실험에서 언급한 가설을 생각해 보면 Aritificial Token은 Attention Mechanism에 의해 높은 가중치를 부여받지 않기 때문에 이와 관련된 정보를 전달하기 위해서는 hidden state와 cell state에 의존할 수 밖에 없고 따라서 정보의 소실을 최소화 하기 위한 방법 이 필요

6. Extras

Input의 순서를 뒤집은 이유

3. 왜 Aritificial Token은 Attention Mechanism에 의해 높은 가중치를 부여받지 않을까?
   - Artificial Token이 담고 있는 어떤 언어로 번역해야 하느냐에 관한 정보는 이전 timestep에 어떤 언어로 번역했는지에 의존하여 현재 timestep에서 어떤 언어의 token을 선택할지를 결정하는 mechanism에서는 Decoding 초기에만 필요로하고 이는 Attention Mechanism이 없어도 Hidden State와 Cell State만 사용해서 충분히 담아낼 수 있기 때문에 Attention Mechanism에서는 이를 중요하다고 여기지 않는듯 함

   - 혹은 Artificial Token이 가지고 있는 정보는 일종의 메타 정보이고 코사인 유사도를 구하는 것으로 Attention Mechanism을 해석한다면 실질적인 의미를 가지는 Token과 이러한 Artificial Token이 완전히 다른 벡터로 표현되어 코사인 유사도 값이 낮게 나와 가중치가 낮아지는 것으로 해석해 볼 수도 있을듯 함.

4. 왜 Artificial Token을 Input의 마지막에 추가하면 안되는 걸까?
   - Artificial Token은 Decoding 초기에만 필요한 정보로 Decoder가 인식하기 때문에 Input의 다른 부분 보다는 초반 부분과 가깝게 두는 게 Decoder가 더 쉽게 번역을 수행할 수 있게 만드는 방법인 듯

Multilingual GNMT와 GPT 비교

Transfer Learning을 통한 다양한 Task가 가능한 모델

두 논문 모두 하나의 모델의 학습으로 다양한 Task를 해결하기 위한 목표를 가지고 있음

GPT-1 논문 인용

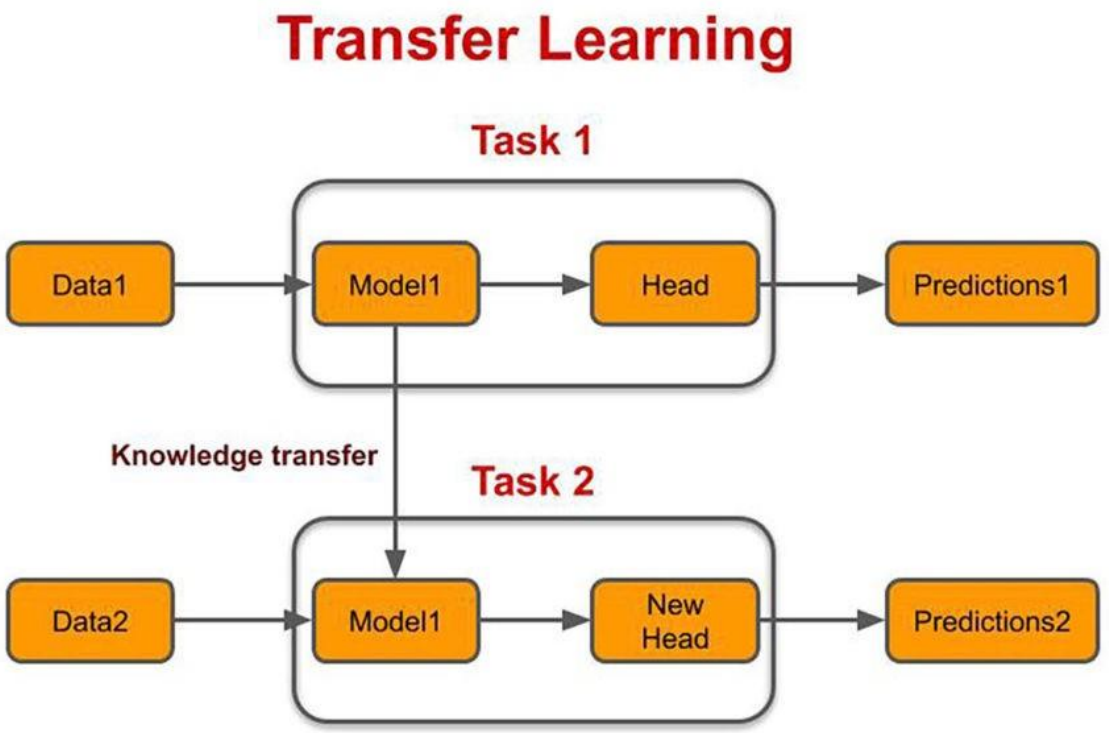
Natural language understanding comprises a wide range of diverse tasks such as textual entailment, question answering, semantic similarity assessment, and document classification.
We demonstrate that large gains on these tasks can be realized by generative pre-training of a language model on a diverse corpus of unlabeled text, followed by discriminative fine-tuning on each specific task.

Multilingual GNMT 논문 인용

We propose a simple solution to use a single Neural Machine Translation (NMT) model to translate between multiple languages.

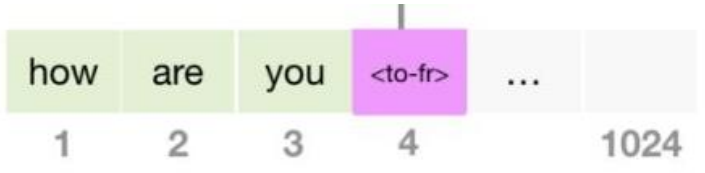
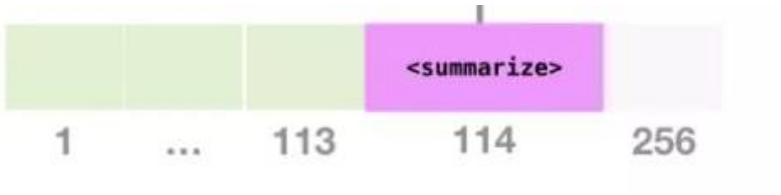
Transfer Learning이란?

Transfer Learning(전이 학습)은 일반적으로 General Task에 대해 훈련 된 모델이 Specialized Task에 어떤 방식 으로든 사용되도록 하는 학습 방법을 말합니다.



Special Token을 사용한 Transfer Learning

GPT

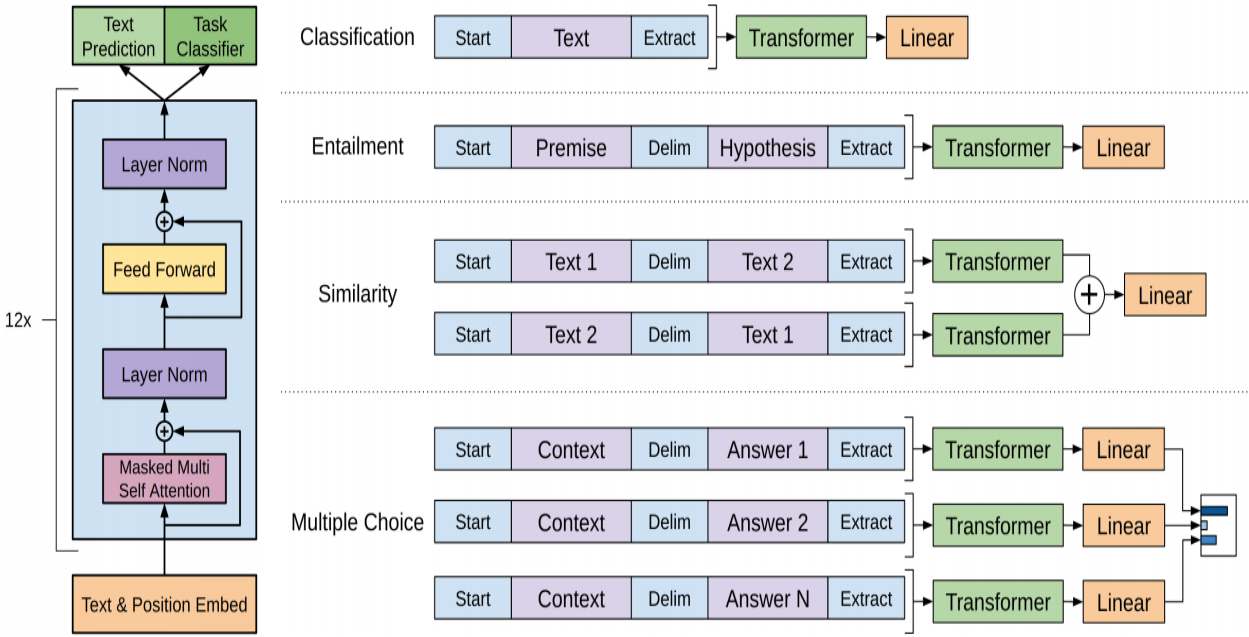
Task	Input
Translation	
Summarize	

Multilingual GNMT

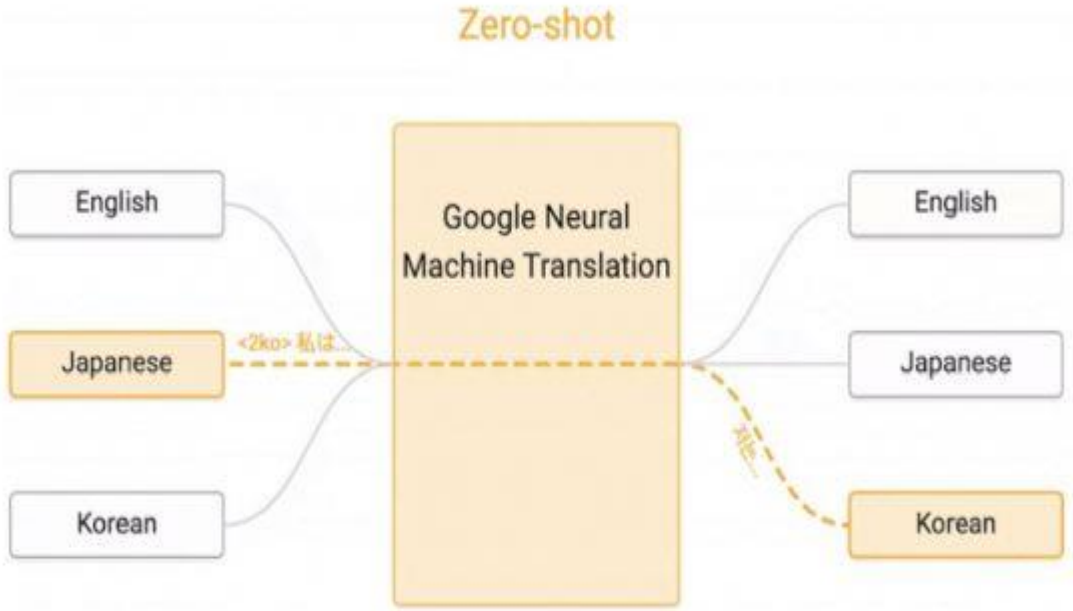
Task	Input
영어->한국어	I am a student <2kr> <eos>
한국어->일본어	나는 학생 입니다 . <2jp> <eos>

다양한 Task가 가능

GPT



Multilingual GNMT



Transfer Learning을 통한 Zero Shot 관점의 차이점

Zero Shot Behaviors VS Zero Shot Translation

	Zero Shot Behaviors	**Zero Shot Translation**
모델	Transformer	LSTM
Zero Shot	• 학습과정에서 수행하지 않은 Task를 Fine Tunning을 통해 가능하도록 함 • 다양한 Task 학습으로 다른 Task 성능 향상	• 학습과정에서 보지 않은 Language pair에 대한 번역을 할 수 있음 • 다른 언어 학습으로 전체적인 번역 성능 향상
예시	**가능한 Task** • 분류 • 요약 • ETC	**학습 Translation** • 영어 -> 한국어 • 한국어 -> 일본어 **가능한 Translation** • 영어 -> 한국어 • 한국어 -> 일본어 • 영어 -> 일본어