

Facebook AI의 WMT21 뉴스 번역 작업 제출

Chau Tran Shruti Bhosale James Cross Philipp Koehn Sergey Edunov 안젤라 팬 Facebook AI
chau,shru,jcross,pkoehn,edunov,angelafan@fb.com

추상적인

뉴스 번역에 대한 WMT2021 공유 작업에 대한 Facebook의 다국어 모델 제출에 대해 설명합니다. 우리는 14개 언어 안내에 참여합니다: 영어와 체코어, 독일어, 하우사어, 아이슬란드어, 일본어, 러시아어, 중국어. 이러한 모든 방향을 포괄하는 시스템을 개발하기 위해 다국어 모델에 중점을 둡니다. WMT, 대규모 데이터 마이닝, 도메인 내 역번역 등 사용 가능한 모든 소스의 데이터를 활용하여 고품질 이중 언어 및 다국어 기준을 만듭니다.

그 후, 하나의 시스템이 8개 언어 모두의 고품질 표현을 위한 충분한 용량을 갖도록 다국어 모델 크기를 확장하기 위한 전략을 조사합니다. 우리의 최종 제출물은 조밀하고 희소한 Mixture-of-Expert 다국어 번역 모델의 앙상블이며, 도메인 뉴스 데이터 및 시끄러운 채널 재순회에서 미세 조정이 뒤따릅니다.

전년도 수상작과 비교했을 때 당사의 다국어 시스템은 모든 언어 방향에서 평균 2.0 BLEU의 번역 품질을 향상시켰습니다. WMT2021 과제에서 우리 시스템은 자동 평가를 기반으로 10개 방향에서 1위를 차지했습니다.

1. 소개

WMT2021 뉴스 번역 공유 과제에 참여하여 다국어 번역 시스템을 제출합니다. 최근 몇 년 동안 다국어 번역은 서로 다른 언어 방향을 위한 별도의 특수 시스템 개발에 대한 대안으로 상당한 관심을 끌었습니다 (Firat et al., 2016; Tan et al., 2019; Aharoni et al., 2019; Zhang et al., 2020; Tang et al., 2020; Arivazhagan et al., 2019).

다국어 시스템은 단순함과 통합에 대한 큰 잠재력을 가지고 있어 매력적입니다. 저희는 체코어(cs), 독일어(de), 하우사어(ha), 아이슬란드어(is), 상업 번역 기술. 다시 일본어(ja), 러시아어(ru), 중국어(zh)에서 온 것입니다. 검색 관련, 전이 학습 연구는 저자원 언어 쌍의 성능을 향상시키는 데 도움이 될 수 있습니다. 이 데이터 소스, 트윈 관련 언어 및 개발 방법을 설명합니다. 205

언어를 하나의 시스템으로 그룹화하기 위한 동기 부여자 (Dabre et al., 2019; Fan et al., 2021).

이러한 동기에도 불구하고 기존의 다국어 번역 시스템은 다국어 시스템의 번역 품질이 이중 언어 시스템을 능가한다는 것을 보여주지 못했습니다. 여러 작업이 이중 언어 기준선과 비교되지만 이러한 기준선은 역번역 또는 조밀한 모델 스케일링과 같이 현장에서 사용되는 기업 표준 기술이 아닙니다.

또한 다국어 번역 시스템은 종종 비표준 교육 데이터 세트에서 개발되고 다른 평가 데이터 세트를 사용합니다. 이러한 요소는 특히 가장 경쟁력 있는 이중 언어 모델과 비교할 때 다국어 번역의 성능을 평가하기 어렵게 만듭니다.

이 작업에서 우리의 목표는 우수한 WMT2020 모델과 이중 언어 WMT2021 시스템에 대해 다국어 번역 모델이 이중 언어보다 더 강력한 성능을 가지고 있음을 입증하는 것입니다. 우리는 14개 언어 방향에 초점을 맞춥니다: 영어와 체코어, 독일어, 하우사어, 아이슬란드어, 일본어, 러시아어, 중국어. 우리는 WMT 분산 및 공개적으로 사용 가능한 교육 데이터를 모두 활용하고, 대규모 역번역을 적용하고, 조밀하고 전문가가 혼합된 아키텍처를 탐색하는 제약 없는 시스템을 만듭니다. 이중 언어 및 다국어 시스템에 대한 다양한 기술의 영향을 비교하여 다국어 시스템이 유리한 부분을 보여줍니다. 최종 다국어 제출은 WMT2020 우승 모델과 비교하여 평균 +2.0의 번역 품질을 개선했으며 WMT2021 리더보드의 자동 평가를 기반으로 10개 방향에서 1위를 차지했습니다.

2 데이터

2.1 바이텍스트 데이터

모든 방향에 대해 공유 작업에서 사용 가능한 모든 바이텍스트 데이터를 사용합니다. 영어에서 독일어 또는 영어에서 러시아어와 같은 언어 지침의 경우 수백만 개의 고품질 바이텍스트를 제공합니다. 그러나 하우스어 및 아이슬란드어와 같은 저-중 자원 언어의 경우 ccMatrix (Schwenk et al., 2019), ccAligned (El Kishky et al., 2020) 및 ccMatrix (Schwenk et al., 2019), OPUS (타이데만, 2012).

사용 가능한 모든 데이터 소스를 활용하여 가능한 최고 품질의 번역 모델을 개발합니다.

English-Hausa(및 Hausa-English)의 경우 제공된 단일 언어 데이터에서 추가 병렬 데이터도 마이닝했습니다. 우리는 LaBSE (Feng et al., 2020) 를 사용하여 하우스어와 영어 문장을 동일한 임베딩 공간에 임베딩 합니다. 그런 다음 KNN(K-Nearest Neighbors)을 기반으로 마진 함수 공식 (Artetxe and Schwenk, 2019) 을 사용하여 두 언어의 문장 쌍에 점수를 매기고 순위를 매깁니다. Tran et al. 의 마이닝 전략 사용 (2020), 우리는 English-Hausa에 대해 백만 쌍의 병렬 문장을 추가로 채웠습니다.

데이터 처리. 사용 가능한 이중 텍스트의 대부분은 인간 번역의 출력 이 아니라 시끄러운 정렬을 나타냅니다. 노이즈 데이터를 필터링하기 위해 여러 단계의 전처리를 적용합니다. 먼저 fasttext (Joulin et al., 2017) 를 사용하여 언어 식별을 적용하고 원하는 언어로 예측된 문장을 유지합니다. 그런 다음 모세를 사용하여 구두점을 정규화합니다. 그런 다음 250단어보다 길고 소스/타겟 길이 비율이 3을 초과하는 문장을 제거했습니다.

2.2 단일 언어 데이터

이전 작업 (Ng et al., 2019) 은 도메인 단일 언어 데이터를 사용하면 대규모 역번역에 사용할 때 가장 품질이 향상된다는 것을 보여줍니다. 영어 및 독일어와 같은 리소스가 많은 언어의 경우 Newscrawl에서 사용할 수 있는 도메인 내 데이터 양이 충분히 많으며 추가 단일 언어 데이터를 사용하지 않습니다.

나머지 언어의 경우 Newscrawl에서 사용할 수 있는 데이터가 충분하지 않으며 다음을 따릅니다.

Moore와 Lewis (2010) 의 전략; Ng et al. (2019) Commoncrawl2에서 대량의 일반 도메인 단일 언어 데이터를 조사합니다.

언어 Bilingual	
체코 어 185M 독일어	140M
571M 하우스어 1.7M 아이슬란드어	237M
어 28.2M 일본어 14M 한국어	700만
어 297M 중국어 106M 영	101M
	218M
	1억5300만
	1억 2300만
—	430M

표 1: 언어당 데이터 양. 바이텍스트 데이터에는 WMT Shared Task, OPUS 저장소, ccMatrix, ccAligned에서 배포한 데이터 및 Hausa에 대해 새로 채굴된 데이터가 포함됩니다. 단일 언어 데이터에는 WMT Shared Task 및 CC100에서 배포한 데이터가 포함됩니다.

사용 가능한 도메인 내 뉴스 데이터와 가장 유사한 하위 집합을 식별합니다. 각 언어에 대해 사용 가능한 모든 뉴스 도메인 데이터(Newscrawl)에 대한 n-gram 언어 모델 (Heafield, 2011) 과 일반 도메인 데이터(Common crawl)의 유사한 크기 샘플에 대한 n-gram 언어 모델을 교육합니다. Commoncrawl의 각 문장에 대해 도메인 내 언어 모델과 일반 도메인 언어 모델을 각각 사용하여 단어 정규화 교차 엔트로피 점수 Hnews(s) 및 Hgeneral(s)을 계산합니다. 임계값 Hnews(s) Hgeneral(s) > 0.01을 충족하는 문장을 유지합니다. 이렇게 하면 원본 Commoncrawl에서 전체 문장 수의 약 5%가 선택됩니다.

2.3 어휘

다국어 어휘를 만들기 위해 먼저 모든 언어에 걸쳐 결합된 교육 데이터에서 다국어 하위 단어 토큰라이저를 배웁니다. 우리는 토큰화되지 않은 텍스트에서 하위 단어 단위를 학습하는 Sentence Piece (Kudo and Richardson, 2018) 를 사용합니다. Conneau et al 과 유사한 온도 업샘플링(T=5 포함)으로 SPM 모델을 훈련합니다. (2020), 저자 원 언어가 표시되도록 합니다. 이중 언어 모델의 경우 32,000개의 어휘 크기를 사용했고 다국어 모델의 경우 128,000을 사용했습니다. 이후 학습된 SPM 단위를 최종 어휘로 변환합니다.

3 시스템 개요

WMT2021에 대한 최종 다국어 제출물을 만든 방법을 단계별로 설명합니다. 이중 언어 및 다국어 모델 아키텍처와 역 번역, 뉴스 도메인 미세 조정, 양상불 및 시끄러운 채널 순위 재 지정 과 같은 전략을 통합하는 방법에 대해 자세히 설명합니다. 206

1참고: 하우스어의 경우 언어 식별 시스템이 불안정하여 사용하지 않았습니다. <http://data.statmt.org/cc-100/>

3.1 기본 이중 언어 모델

최첨단 다국어 번역 시스템을 만들기 위한 전제 조건은 강력하고 경쟁력 있는 이중 언어 기준을 설정하는 것입니다. 우리의 목표는 이중 언어 및 다국어 모델 모두에 데이터 확대 및 모델링 스케일링에서 동일한 기술 세트를 적용하고 다국어 모델이 더 강력한 번역 품질을 가지고 있음을 입증하는 것입니다.

기본 이중 언어 시스템을 만들기 위해 각 언어 방향에 대해 별도의 Transformer 모델 (Vaswani et al., 2017) 을 교육합니다. Hausa를 제외한 모든 언어 쌍에 대해 표 2의 Transformer 12/12 구성을 사용합니다. Hausa-English(및 English-Hausa)의 경우 바이텍스트 데이터의 양이 적기 때문에 Vaswani et al.과 유사한 Transformer-Base 아키텍처를 사용합니다. 알. (2017). 우리는 32개의 Volta 32GB GPU에서 fairseq (Ott et al., 2019) 를 사용하여 모든 모델을 훈련합니다. 우리는 Adam 옵티마이저와 함께 0.001의 학습률을 사용하고 kens3 방향에 대해 768,000의 배치 크기를 독립적으로 사용합니다. 대형 모델용 , 각 언어의 탈락률을 조정합니다.

3.2 역번역

역번역 (Sennrich et al., 2015) 은 데이터 증대를 사용하여 기계 번역 시스템의 품질을 개선하기 위해 널리 사용되는 기술입니다. 순방향 언어 방향(예: 영어에서 독일어로)에 대한 역번역을 수행하기 위해 우리는 대상 독일어 단일 언어 데이터를 영어 소스로 번역하기 위해 역방향(예: 독일어에서 영어로) 시스템을 사용합니다. 그런 다음 개선된 정방향 번역 모델을 교육하기 위해 원본 병렬 데이터와 함께 역번역된 합성 영어에서 독일어 문장 쌍을 사용합니다.

우리는 역번역을 위해 각 언어(언어당 최대 5억 문장)에 대해 사용 가능한 모든 필터링된 단일 언어 데이터를 사용합니다. 기본 이중 언어 모델(섹션 3.1에서 설명)을 사용하여 먼저 도메인 내 뉴스 데이터 (섹션 3.5에서 설명) 를 미세 조정 하고 다른 시드가 있는 3개 모델의 앙상블을 사용하여 빔 검색을 사용하여 역번역 데이터를 생성합니다. Hausa-English 및 English-Hausa의 경우 품질 개선이 중요하므로 반복적인 역번역 (Hoang et al., 2018; Chen et al., 2019) 을 적용했습니다.

3.3 데이터 샤딩 및 샘플링

표 1은 후처리 후 모든 언어에 대한 데이터 양을 표시합니다. 우리는 데이터를 나눕니다.

	12/12	24/24	24/24	와이드
레이어	12	24	1,024	24
Emb. 크기	1,024	4,096	8,192	2,048
FFN 크기 Attn.	16	16		16,384
헤드				32
총 매개변수. 480M	1.2B			4.7B

표 2: 고밀도 변압기 구성.

각 훈련 에폭이 하나의 샤드를 사용하는 여러 개의 샤드로. 우리는 높은 리소스 방향과 합성 역변환 데이터 모두에서 데이터를 낮은 리소스 방향의 실제 바이텍스트 데이터보다 더 많은 수의 샤드로 나누어 다운샘플링합니다. 업샘플링이 과적합에 더 크게 기여하기 때문에 다운샘플링 고자원 언어가 저자원 언어를 업샘플링하는 것보다 더 잘 작동한다는 것을 알았습니다.

3.4 모델 아키텍처

이중 텍스트 및 역번역 데이터가 있는 최종 데이터 세트를 사용하여 비교한 여러 모델 아키텍처를 설명합니다.

이중 언어 모델 확장. 섹션 3.1에 설명된 기본 아키텍처를 기반으로 이중 언어 모델을 더욱 개선합니다. 두 가지 주요 개선 사항은 역변환된 데이터를 추가하는 것과 데이터 증가를 활용하기 위해 더 깊고 더 넓은 Transformer 구성을 추가하는 것입니다.

조밀한 다국어 모델. 다국어 시스템의 경우 두 가지 개별 모델을 교육합니다. Many to English 또는 모든 언어를 영어로 번역하는 하나의 시스템과 English to Many 또는 영어를 모든 언어로 변환하는 시스템입니다. 다국어 모델의 과제는 종종 용량입니다. 고정된 수의 매개변수가 주어지면 모델은 하나가 아닌 여러 언어의 표현을 학습해야 합니다. 필요한 용량과 최적의 아키텍처 구성을 이해하기 위해 480M 매개변수에서 4.7B 매개변수에 이르는 다양한 Transformer 아키텍처를 실험했습니다 (표 2 참조).

성가게 게이트된 MoE 다국어 모델. 다국어 모델에서 언어는 반드시 용량을 놓고 경쟁하며 다른 언어에 대한 전문화와 공유 매개변수의 균형을 유지해야 합니다. 신경 구조에 용량을 추가하는 간단한 방법은 모델 크기를 조밀한 방식으로 확장하는 것입니다. 즉, 레이어 수, 레이어 너비 또는 숨겨진 차원의 크기를 늘리는 것입니다. 207

[†] GPU당 토큰 6000개 * GPU 32개 * 업데이트 빈도 4개

그러나 이것은 각 순방향 패스가 모든 매개변수를 활성화하기 때문에 상당한 계산 비용이 듭니다. 한계에서 모델은 번역을 학습하고 생성하는 데 엄청나게 느려집니다 (Fan et al., 2021).

이 작업에서는 계산 비용의 비례적 증가 없이 용량을 늘리고 자 동기가 부여된 스파스 모델 스케일링에 초점을 맞춥니다. 우리는 Sparsely Gated Mixture-of Expert(MoE) 모델 (Lepikhin et al., 2020) 을 English to Many 및 Many to English 로 훈련합니다. 이러한 모델은 공유 모델 용량을 통해 리소스가 적은 방향으로 전환하는 동시에 전문 모델 용량 증가의 이점을 얻기 위해 높은 리소스 방향을 허용하는 것 사이에서 균형을 맞추는 것을 목표로 합니다.

각 Sparsely Gated MoE 레이어에서 각 토큰은 학습된 게이팅 기능을 기반으로 상위 k 전문가 FFN 블록으로 라우팅됩니다. 따라서 입력 시퀀스당 모든 모델 매개변수의 하위 집합만 사용됩니다.

Feed와 함께 Transformer 아키텍처를 사용합니다.

인코더 및 디코더에서 상위 2개 게이팅이 있는 Sparsely Gated Mixture-of-Experts 레이어로 대체된 모든 대체 Transformer 레이어의 전방 블록.

Lepikhin et al. (2020), 게이트 손실 가중치가 0.01인 토큰에서 전문가 할당의 균형을 맞추기 위해 게이트 손실 기간도 추가합니다. 전문가 용량 계수 2.0을 사용합니다. 우리는 4000개의 위밍업 업데이트와 100만 토큰(64명의 전문가가 있는 MoE 모델) 또는 150만 토큰(128명의 전문가가 있는 MoE 모델)의 배치 크기가 있는 Adam 옵티마이저와 함께 0.001의 학습률을 사용합니다.

3.5 도메인 내 미세 조정

도메인별 데이터로 미세 조정하는 것은 원하는 도메인에 대한 번역 품질을 향상시키는 효과적인 방법입니다. 미세 조정을 위해 뉴스 도메인 데이터를 큐레이팅했습니다. 독일어 및 러시아어와 같은 방향의 경우 이전 WMT의 평가 데이터 세트를 미세 조정합니다. 하우스어 및 아이슬란드어의 경우 이전 데이터가 없으므로 마이닝된 데이터를 사용하고 가장 가능성이 높은 뉴스 도메인으로 식별된 하위 집합으로 필터링합니다. 그 후, 우리는 맥시를 위해 도메인 내 데이터에서 모델을 미세 조정합니다.

newstest2020 개발 세트에서 유효성 검사 손실이 있는 최상의 모델을 선택하는 10개의 에포크의 엄마입니다.

제출을 위해 newstest2020에 조정된 설정을 사용하고 newstest2021을 포함합니다.

최종 미세 조정에서 dev 세트.

3.6 체크포인트 평균화

최근 훈련 데이터에 대한 편향을 방지하기 위해 모델의 여러 체크포인트에서 매개변수를 평균화하는 것이 일반적입니다 (Vaswani et al., 2017). 우리 208

이 기술을 모든 모델에 적용하고 마지막 5개 체크포인트의 평균을 냅니다. 미세 조정 중 빠른 과적합 문제를 해결하기 위해 초기 교육이 완료된 후 모델로 미세 조정된 모델의 평균을 내고 검증 데이터에서 더 나은 성능을 보이는 경우 이 평균 매개 변수 집합을 선택합니다.

3.7 시끄러운 채널 재순위

Beam search로 생성된 n -best 가설에서 최고의 후보 번역을 선택하기 위해 noise channel re-ranking을 적용합니다.

Yee 등을 따릅니다.

(2019); Bhosale et al. (2020) 및 직접 모델 $P(\text{tgt}|\text{src})$, 채널 모델 $P(\text{src}|\text{tgt})$ 및 언어 모델 $P(\text{tgt})$ 의 점수를 활용합니다. 순위를 재지정하기 위해 이러한 점수를 결합하기 위해 모든 n -best 가설에 대해 다음을 계산합니다.

$$\log P(\text{tgt}|\text{src}) + \lambda_1 \log P(\text{src}|\text{tgt}) + \lambda_2 \log P(\text{tgt})$$

가중치 λ_1 및 λ_2 는 1000회 이상의 임의의 검색으로 조정하여 결정됩니다.

검증 세트를 선택하고 최상의 성능을 제공하는 가중치를 선택합니다. 또한 길이 패널티도 조정합니다. 가중치와 길이 패널티에 사용하는 검색 범위는 $[0, 2]$ 입니다.

언어 모델. 우리는 역번역에 사용된 것과 동일한 단일 언어 데이터로 모든 언어에 대한 Transformer 기반 언어 모델을 교육했습니다. CC100 영어 데이터와 RoBERTa 훈련 데이터로 훈련한 영어는 예외입니다 (Conneau et al., 2020; Wenzek et al., 2019; Liu et al., 2019). 리소스가 많은 언어의 경우 언어 모델에는 12개의 디코더 레이어와 임베딩 차원 4096이 있습니다. Hausa 및 Icelandic의 경우 오버피팅을 방지하기 위해 6개의 디코더 레이어로 더 작은 언어 모델을 교육했습니다.

3.8 후처리

마지막 단계로 체코어, 독일어, 아이슬란드어, 일본어 및 중국어 번역 출력에 사후 처리를 적용합니다. 체코어, 독일어 및 아이슬란드어의 경우 따옴표를 독일어 큰따옴표 style4로 변환합니다.

중국어와 일본어의 경우 구두점을 언어별 구두점 문자로 변환합니다.

4 실험 및 결과

우리는 시스템의 각 구성 요소가 미치는 영향을 정량화하기 위해 실험을 수행합니다. 모든 실험은 SacreBLEU (Post, 2018)를 사용하여 newstest20 (Barrault et al., 2020)에서 평가되었습니다.

⁴https://en.wikipedia.org/wiki/Quotation_mark#독일어

	cs-en	de-en	ha-en	is-en	ja-en	ru-en	zh-en
다국어 단어 28.3 이중 언어 단어 28.6	38.0 36.8	28.3 26.8	21.1 20.4	28.4 26.1	35.2 33.7	22.4 21.9	38.0 37.0
	en-cs	en-de	en-ha	en-is	en-ja	en-ru	en-zh
다국어 단어 33.2 이중 언어 단어 33.7	39.4 39.8	23.1 29.4	29.4 26.1	23.9 26.1	25.7 26.0	42.4 43.3	

표 3: 이중 언어 모델에 대한 어휘의 영향. 특수 이중 언어 어휘와 일반 다국어 어휘를 사용하여 이중 언어 시스템의 성능에 미치는 영향을 비교합니다.

4.1 최첨단 다국어 생성 번역 모델

우리는 번역에서 다국어의 효과를 조사합니다. 특정 소스 및 대상 언어에 전문화할 수 있는 이중 언어 모델과 비교할 때 다중 언어 시스템은 다양한 리소스 수준의 언어 균형을 유지하면서 모든 언어에서 사용 가능한 용량을 효과적으로 공유하는 방법을 배워야 합니다. 증가하는 연구 관심에도 불구하고 이전 WMT 제출은 이중 언어 모델보다 다국어 모델의 품질 향상을 입증하지 못했습니다. 최첨단 다국어 번역 시스템을 구축하는 다양한 선택과 비교에 대해 논의합니다. 전반적으로 최고의 다국어 시스템은 14개 방향 중 11개 방향에서 최고의 이중 언어 시스템을 능가하며 평균 +0.8 BLEU 향상됩니다.

4.1.1 다국어 어휘 구축.

다국어 시스템이 모델 용량을 공유해야 하는 방식과 유사하게 다국어 번역 모델도 어휘 용량을 공유해야 합니다. 특정 언어(종종 32k)에 대한 특수 하위 단어 단위를 훈련하는 대신 다국어 모델은 모든 언어를 함께 그룹화하여 32k * 언어 수보다 훨씬 작은 vocabulary 세트를 학습합니다. 우리는 먼저 이중 언어 시스템을 선택하고 다중 언어 어휘로 훈련함으로써 이 다중 언어 어휘의 영향을 조사합니다. 이것은 아키텍처가 아니라 어휘 자체에서 오는 성능 차이를 나타냅니다. 표 3은 모든 방향에서 특수 이중 언어 어휘를 사용하는 것이 일반적으로 우수하다는 것을 나타냅니다. 즉, 다국어 시스템은 잠재적으로 수준 이하인 어휘의 성능 격차를 메워야 합니다. 그러나 en-is 및 en-ja와 같은 일부 방향에서는 차이가 관찰되지 않습니다.

4.1.2 모델 아키텍처 비교.

전반적으로 밀도가 높은 다국어 모델이 보드와 상당히 경쟁하는 성능을 보여줍니다. 관련 209와 비교하여 MoE 모델의 조정

중요한 점은 다국어 모델이 추가 모델 용량에서 큰 이점을 얻는다는 것입니다. 표 5에서 우리는 영어에서 독일어로 번역하는 이중 언어 모델에 적용된 유사한 고밀도 스케일링을 보여줍니다. 다국어 모델은 최대 1 BLEU 포인트까지 향상되지만 이중 언어 모델은 +0.3 BLEU 만 향상되어 이중 언어 모델에서 수익 감소 및 과적합 가능성이 있음을 나타냅니다. 다국어 번역 모델을 확장하면 성능 향상 가능성이 더 커집니다.

전문가 모델의 희소하게 게이팅된 혼합. 다국어 모델이 더 큰 용량의 혜택을 받는 경우 해당 용량을 추가하는 가장 좋은 방법은 무엇입니까? 표 4에서는 공정한 비교를 위해 업데이트당 FLOP를 거의 동일하게 유지하면서 Dense 및 MoE 다국어 모델의 성능을 비교합니다. MoE 레이어의 조건부 계산 용량으로 인해 MoE 모델은 더 많은 총 매개변수를 갖지만 해당 조밀한 모델과 비슷한 계산 비용을 갖습니다.

Many to English 및 English to Many의 경우 MoE 계층당 64명의 전문가가 있는 MoE 모델은 개발 세트에서 +0.7 BLEU의 평균 향상을 제공합니다.

밀집 모델 스케일링과 비교하기 위해 밀집 모델 크기를 12/12에서 24/24로 늘리는 것은 그렇지 않습니다.

Many to English의 상당한 개선에 해당합니다. 그러나 English to Many의 고밀도 확장에서 약 +1 BLEU 개선이 있습니다.

또한 모델 차원을 늘리고 전문가 수를 64에서 128로 늘릴 때 MoE 모델(MoE-64 12/12 대 MoE-128 24/24)의 성능이 약간 감소하거나 개선되지 않는 것을 볼 수 있습니다. 한 가지 가능한 가설은 다음과 같습니다. 단 7개 언어에 대해 128명의 전문가가 필요하지 않다는 것입니다. 64명의 전문가에 비해 평균적으로 각 전문가가 교육 중에 더 적은 토큰에 노출되기 때문에 전문가당 교육 수렴이 느립니다.

도메인 내 데이터를 미세 조정 후 고밀도 변압기 모델을 관찰합니다. 전문가는 사용 가능한 용량에 비해 향상된 성능을 보여줍니다. 관련 209와 비교하여 MoE 모델의 조정

	cs-en de-en ha-en is-en ja-en ru-en zh-en 평균									
이중 언어 밀집 12/12	28.3	38.0	28.3	34.5	21.1	28.3	35.2	38.0	30.8	31.3
밀집 12/12	26.9	37.5	19.0	27.2	37.3	18.5		36.2	28.8	30.3
MoE-64 12/12	28.0	38.9						39.1	28.0	31.0
조밀한 24/24	28.1	37.2	26.3	35.6	20.6	36.9		35.8	28.0	30.2
MoE-128 24/24	28.1	36.8	23.1	18.7	36.8	21.2		36.9	29.7	29.7
조밀한 24/24 와이드	29.0	37.9	24.5					36.9	30.4	31.0
이중 언어 밀집 12/12, BL-FT 30.4 밀집 12/12, ML-FT 30.3 MoE-64 12/12, ML-FT 31.6	42.8	30.3	35.5	24.6	39.5	36.2		34.2		34.2
	42.4	32.7	37.5	23.9	39.5	34.2		34.2		34.4
	43.5	33.4	38.8	25.7	39.8	36.0		36.0		35.5
고밀도 24/24, ML-FT	31.8	43.4	36.0	38.8	25.6	34.9	39.7	40.3	36.3	36.0
MoE-128 24/24, ML-FT	31.9	43.6	26.5	36.1	39.4	26.7		40.4	37.2	36.3
조밀한 24/24 와이드, ML-FT	32.1	43.8						40.6	36.9	36.5
	en-cs en-de en-ha en-is en-ja en-ru en-zh Avg									
이중 언어 밀집 12/12	33.1	39.6	23.1	29.4	26.1	30.5	25.7	42.4		31.3
밀집 12/12	33.7	38.6	21.4	26.6	31.5	28.0	25.3	41.1		31.0
MoE-64 12/12	33.5	39.7	20.4				26.4	42.5		31.7
조밀한 24/24	34.0	39.6	21.7	31.6	27.5	19.3	30.9	26.4	42.3	31.9
MoE-128 24/24	33.0	40.2	28.8	23.4	32.0	28.0		26.6	42.8	31.7
조밀한 24/24 와이드	33.4	39.7						26.6	42.2	32.2
이중 언어 밀도 12/12, BL-FT 35.7 밀도 12/12, ML-FT 35.0 MoE-64 12/12, ML-FT 35.9	39.5	23.3	29.4	27.7	30.5	26.0	43.0			32.1
	39.1	22.9	26.9	29.6	28.8	25.6	41.5			31.6
	40.4	24.1				26.4	43.0			32.6
고밀도 24/24, ML-FT	35.8	40.1	24.1	31.6	28.7	24.6	31.2	26.8	42.5	32.8
MoE-128 24/24, ML-FT	36.4	40.8	29.7	24.6	32.0	29.3		26.8	43.6	33.3
조밀한 24/24 와이드, ML-FT	36.7	40.6						26.7	43.0	33.3

표 4: 영역 내 미세 조정 전과 후의 조밀한 대 희소하게 게이트된 MoE 다국어 모델 비교 . BL-FT는 이중 언어 데이터에 대한 모델 미세 조정을 의미하고 ML-FT는 다국어 데이터에 대한 모델 미세 조정을 의미합니다(섹션 4.1 참조).

	양데
이중 언어 12/12	39.8
이중 언어 24/24	40.1
이중 언어 24/24 와이드	40.3
이중 언어 12/12 + FT 40.4 이중 언어 24/24 + FT 40.5 이중 언어 24/24 와이드 + FT 40.4	

표 5: 이중 언어 모델 확장.

조밀한 기준선. 또한 128명의 전문가를 가진 MoE 모델은 이전에 64명의 전문가를 가진 MoE 모델보다 뒤떨어져 이제 두 가지 방향을 제외한 모든 방향에서 최상의 결과를 제공합니다 . 가능한 가설은 MoE 모델의 전문가 역량이 미세 조정하는 동안 모든 언어 방향이 모든 모델 용량을 공유해야 하는 조밀한 모델보다 특수한 방향별 미세 조정을 더 잘 유지할 수 있다는 것입니다 .

4.1.3 도메인 내 미세조정 효과 미세조정은 이중언어

보다 다국어를 향상시킨다. 표 6 은 미세 조정의 영향을 비교한 것입니다. 다양한 모델에 대한 튜닝이 여러 차례 수행되었습니다. BLEU cor 시스템이 도메인 내 미세 조정에서 더 많은 이점을 얻는 방법을 이해하기 위한 이중 언어 모델의 다국어 . 인간이 인지하는 품질과 관련됨, 소수의 연구 210

결과적으로 최고의 다국어 시스템은 항상 최고의 이중 언어 시스템을 능가합니다.

다중 언어 미세 조정은 이중 언어 미세 조정보다 낮습니다. 다국어 모델의 경우 두 가지 미세 조정 방식이 있습니다 (Tang et al., 2020). 다국어 모델은 모든 언어에 대한 뉴스 데이터를 연결하여 다국어 방식으로 뉴스 도메인에 특화되도록 미세 조정하거나 이중 언어 뉴스 도메인 데이터에 대한 교육을 통해 각 방향에 대해 개별적으로 미세 조정할 수 있습니다. 표 6에서 다국어 도메인 내 미세 조정 과 이중 언어 도메인 내 미세 조정을 비교합니다 .

우리는 다국어 미세 조정이 이중 언어 미세 조정보다 거의 모든 면에서 낫다는 것을 발견했습니다. 즉, 다국어 시스템을 가져와 이중 언어 미세 조정을 통해 이중 언어가 되도록 전문화할 필요가 없음을 나타냅니다. 완전한 다국어 시스템이 여전히 가장 강력한 성능을 가 집니다 .

4.1.4 인간 평가.

	cs-en de-en ha-en is-en ja-en ru-en zh-en							
이중 언어	28.3	38.0	30.4	28.3	34.5	21.1	38.0	30.8
이중 언어, BL-FT	42.8			30.3	35.5	24.6	39.5	36.2
다국어 29.0 다국어, BL-FT 31.8 다국어, ML-FT 32.1		37.9		24.5	36.8	21.2	31.9	37.0
		43.3		26.5	36.1	39.4	26.7	40.6
		43.8					40.6	36.9

	en-cs en-de en-ha en-is en-ja en-ru en-zh							
이중 언어	33.1	39.6	23.1	29.4	26.1	23.3	29.4	25.7
이중 언어, BL-FT	35.7	39.5	27.7					26.0
다국어 33.4 다국어, BL-FT 36.1 다국어, ML-FT 36.7		39.7	23.4	32.0	28.0	24.2	30.1	26.6
		40.3	28.7	24.6	32.0	29.3		27.4
		40.6						26.7

표 6: 이중 언어 및 다국어 모델에 대한 미세 조정의 영향. BL-FT는 이중 언어 데이터에 대한 다국어 모델 미세 조정을 의미하고 ML-FT는 다국어 데이터에 대한 다국어 모델 미세 조정을 의미합니다.

cs-en de-en ha-en is-en ja-en							
이중 언어 28.9 + BT	41.5	30.3	18.5	34.5	38.0	21.1	-3.5
28.3 Δ -0.6	+1.4		28.3		+12.4	+4.2	

en-cs en-de en-ha en-is en-ja							
이중 언어 33.1 + BT	38.7	14.7	25.8	25.4	23.1	29.4	
33.2 Δ +0.1	39.4	26.1	+8.4	+3.6	+0.7		
	+0.7						

표 7: 이중 언어 시스템에서 대규모 역번역의 영향.

cs-en de-en ha-en is-en ja-en							
다국어 27.7 + BT 27.8 Δ	37.6	16.5	34.2	20.8	25.8	35.6	
+0.1	37.9	20.8	+9.3	+1.4	+0		
	+0.3						

en-cs en-de en-ha en-is en-ja							
다국어 33.7 + BT 33.9 Δ	39	10.0	27.0	26.9	31.6		
+0.2	39.2	23.7	27.6	+13.7	4.6		
	0.2	+0.7					

표 8: 다국어 시스템에서 대규모 역번역의 영향.

결과에 따르면 동일한 BLEU 점수로 훈련된 시스템이 대규모 역번역에 대한 성능이 향상되어 다국어 시스템과 비교하여 다른 언어로 인간의 평가에 영향을 미칠 것입니다. 번역할 때 낮은 품질의 번역을 유도하는 것은 큰 단점입니다 (Maru et al., 2020). 영어에서 영어로 이해로, 우리는 인간 어노테이터에게 역번역이 근본적으로 소스에 능통하고 대상 언어에 비해 더 나은 번역을 생성할 수 있도록 하는 데 도움이 될 수 있습니다. 다국어 시스템 출력은 교육 데이터 세트에 음바를 통합하고 있습니다. Beyond에서 다국어 번역도 소수점 차이를 볼 수 있습니다. 번역은 동일한 영어 소스 문장에 입력된 데이터 하집의 균형을 형성합니다.

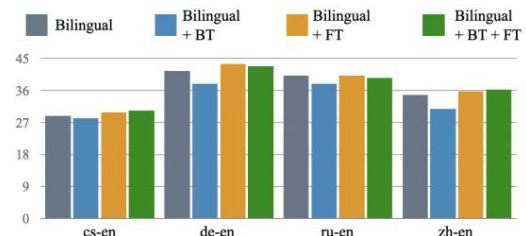


그림 1: 역번역 후 도메인 내 미세 조정이 이중 언어 모델에 미치는 영향.

인간 평가자가 이중 언어 또는 다중 언어 시스템을 선천적으로 선호하지 않음을 나타내는 두 시스템의 인간 평가 사이에 통계적으로 유의미한 차이가 없음을 발견했습니다.

4.2 대규모 역번역의 영향

대규모 역번역은 인간 평가 연구 (Edunov et al., 2019; Bogoychev and Sennrich, 2019) 에서 측정된 경우에도 기계 번역 모델 (Edunov et al., 2018)의 성능 향상에 기여했습니다. 대부분의 현대 번역 시스템에 통합된 Nent는 다국어 번역 시스템을 조사했습니다. 빌린 tems가 주어집니다. 그러나 역번역에는 단점도 있습니다. guual 시스템과 다국어 시스템 연구에 따르면, 다국어 시스템은 다국어 문장과 비특정적으로 다른 언어로 인간의 평가에 영향을 미칠 것입니다. 번역할 때 낮은 품질의 번역을 유도하는 것은 큰 단점입니다 (Maru et al., 2020). 영어에서 영어로 이해로, 우리는 인간 어노테이터에게 역번역이 근본적으로 소스에 능통하고 대상 언어에 비해 더 나은 번역을 생성할 수 있도록 하는 데 도움이 될 수 있습니다. 다국어 시스템 출력은 교육 데이터 세트에 음바를 통합하고 있습니다. Beyond에서 다국어 번역도 소수점 차이를 볼 수 있습니다. 번역은 동일한 영어 소스 문장에 입력된 데이터 하집의 균형을 형성합니다.

MMT 모델		cs-en de-en ha-en is-en ja-en ru-en zh-en 평균 증분 Δ									
이중언어 정 30.4	28.9 + 역번역 28.3 + 미세조	41.5	15.9	30.3	19.7	40.2	34.8	30.2	—		
		38.0	28.3	34.5	21.1	38.0	30.8	31.3	+1.1		
		42.8	30.3	35.5	24.6	39.5	36.2	34.2	+2.9		
✓ + 다국어 킹	✓ + 앙상블 ✓ + 리랭	32.1	43.8	36.1	39.4	26.7	39.9	40.6	36.5	+2.3	
		32.3	44.5	37.2	27.2	40.5	27.8	40.9	37.8	37.1	+0.6
		32.7	44.4	38.2				41.4	38.0	37.6	+0.5
WMT20 승자 29.9 39.2 WMT20보다 Δ 2.8 +0.6 -43.8 ± 2.2								36.9			
								+1.1			

MMT 모델		en-cs en-de en-ha en-is en-ja en-ru en-zh 평균 증분 Δ									
이중언어 정 35.7	33.1 + 역번역 33.1 + 미세조	38.7	14.7	25.8	25.4	23.1	29.4	25.8	40.0	29.1	—
		39.6	26.1	23.3	29.4	27.7		25.7	42.4	31.3	+2.3
		39.5						26.0	43.0	32.1	+0.7
✓ + 다국어 재지정	✓ + 앙상블 ✓ + 순위 ✓ + 후처리	36.4	40.8	24.6	31.2	29.7	32.5	26.8	43.6	33.3	+1.2
		36.8	41.1	25.0	29.7	32.8	29.7	26.9	43.6	33.7	+0.4
		37.2	41.1	25.5	29.8			27.4	43.6	33.9	+0.2
		39.8	42.6	25.5				28.8	48.2	35.6	+1.7
WMT20 우승자 36.8 25.5 WMT20보다 Δ 3.0 +3.8 38.8 -1.4 28.8								47.3			
								+0.9			

표 9: 제출된 모델의 전체 결과. 이중 언어 기준으로 시작하여 언어 쌍에 걸쳐 다양한 기술의 증분 이득을 설명합니다. 최종 제출물은 역번역을 포함하여 사용 가능한 모든 데이터에 대해 훈련된 시끄러운 채널 재순위가 있는 다국어 앙상블입니다. 모든 언어 쌍에서 이전 WMT20 수상 모델과 비교하여 개선이 관찰됩니다. MMT 열은 모델이 다국어인지 여부를 나타냅니다. 참고 하우사어와 아이슬란드어는 WMT20에 없었습니다.

다른 방향의. 따라서 이 섹션에서는 다국어 시스템에서도 역번역의 지속적인 중요성에 대해 자세히 분석합니다.

이중 언어 시스템의 역번역. 먼저 WMT 공유 작업에 배포된 것으로 마이닝되고 공개적으로 사용 가능한 훈련 데이터로 훈련 데이터 세트를 보강한 후에도 역번역 데이터가 여전히 유용한지 조사합니다. 표 7의 결과는 역변환이 14개 방향 중 10개 방향, 특히 ha-en 및 is-en과 같은 리소스가 적은 방향에 도움이 된다는 것을 보여줍니다. 그러나 de-en, ru-en, zh-en과 같은 높은 리소스 방향의 경우 역번역으로 훈련된 이중 언어 시스템은 역번역 없이 훈련된 시스템에 비해 유효성 BLEU가 약간 낮았습니다.

Finetuning은 Translationese에 대한 Overfitting을 수정합니다. 높은 리소스 방향이 역번역된 데이터를 추가할 때 어려움을 겪을 수 있는 이상 현상을 추가로 조사합니다. 그림 1은 역번역 추가로 인한 mi 또는 BLEU 저하가 도메인 내 미세 조정을 적용한 후 대부분 사라짐을 보여줍니다. 도메인 내 미세 조정 후 zh-en 및 cs-en의 경우 역번역으로 훈련된 시스템이 역번역 없이 훈련된 시스템에 비해 더 강력한 성능(+0.4 BLEU)을 가집니다. 이 효과에 대한 이전 연구는

그 역번역은 원본 학습 데이터와 비교하여 뚜렷한 품질을 가진 번역어를 생성합니다 (Marie et al., 2020; Zhang and Toral, 2019; Graham et al., 2020). 우리는 역번역되지 않은 데이터에 대해 모델을 훈련시키는 도메인 미세 조정에서 번역어에 대한 과적합에 대응하는 수정 효과를 가질 수 있다고 가정합니다.

다국어 시스템의 역번역. 표 8은 절제 연구에서 다국어 모델에 역번역을 추가하여 성능 향상을 요약한 것입니다. 전반적으로 훨씬 더 많은 훈련 데이터로 완전히 제한되지 않은 시스템을 만들고 다국어 번역의 데이터 공유 잠재력을 활용함에도 불구하고 우리는 역번역이 여전히 성능을 향상시킨다는 것을 발견했습니다.

이는 역번역이 사용 가능한 단일 언어 데이터를 완전히 활용한다는 사실에 의해 영향을 받는 것으로 생각됩니다. 데이터 마이닝 기술은 잠재적으로 병렬 문장을 식별할 수 있지만 알고리즘이 마이닝에 활용하는 전체 단일 언어 데이터의 하위 집합만 식별하는 것으로 자연스럽게 제한됩니다.

4.3 최종 부품에 대한 절제 제출

마지막으로 최종 제출물의 각 측면과 누적 효과를 분석하여 마무리합니다. 그만큼

각 성분의 효과는 표 9와 같다. 212

이중 언어 기준. 우리는 이중 언어 기준선이 특히 WMT20 승자보다 이미 이중 언어 기준선이 더 강한 ru-en의 경우 BLEU 점수가 높다는 것을 발견했습니다. 전반적으로 우리는 en-ha와 ha-en만이 20 BLEU보다 현저히 낮음을 관찰했으며, 이는 대량의 고품질 바이텍스트 데이터를 큐레이팅하는 것이 강력한 시스템의 가장 중요한 기반일 가능성이 있음을 나타냅니다.

역번역. 그런 다음 번역된 데이터를 다시 추가합니다. 특히 ha, is 및 ja는 역번역된 데이터를 추가한 후 BLEU에서 큰 개선을 관찰하는 반면 다른 방향은 번역어의 가능한 영향으로 실제로 품질이 약간 저하될 수 있습니다.

도메인 내 미세 조정. 다음으로 도메인 내 미세 조정의 영향을 평가하고 영어로 번역하는 방향에 대해 거의 3 BLEU 개선, 영어 외 번역에 대해 0.7 BLEU 개선을 찾습니다. 모든 언어 방향에서 미세 조정은 거의 보편적으로 도움이 됩니다.

다국어. 이중 언어 모델에 비해 다국어 모델은 모든 방향에서 더 강력한 성능을 보입니다. 우리의 가장 큰 아키텍처 (MoE-128 24/24)가 최고의 성능을 발휘하므로 다국어 모델은 모델 크기를 확장하면 훨씬 더 많은 이점을 얻을 수 있습니다.

양상블. 양상블이 평균에 미치는 영향은 상당히 미미하지만 특정 방향에서는 큰 개선을 볼 수 있습니다(예: zh-en에서 +1 BLEU).

재순위. 그런 다음 최종 시스템의 출력에 노이즈 채널 재순위를 적용합니다. 거의 모든 방향에서 유용 하지만 BLEU에는 큰 영향을 미치지 않습니다. 평균적으로 성능은 약 0.3~0.5 BLEU 향상됩니다.

사후 처리. 마지막으로, 각 언어에서 표준화된 구두점을 사용하기 위해 번역된 출력을 사후 처리하는 것이 영어 이외의 언어로 번역할 때 BLEU 점수에 매우 중요하다는 것을 관찰했습니다. 예를 들어, 특히 중국어에는 많은 특정 마침표와 이중 폭 문장 부호 문자가 있으며 이를 적절하게 사용하면 거의 +5 BLEU가 생성됩니다. 그러나 이러한 기술은 BLEU 점수만 향상시킬 가능성이 높으며 인간 평가에 미치는 영향은 잘 알려져 있지 않습니다.

5. 결론

이 백서에서는 WMT2021 공유 작업에 대한 Facebook의 다국어 모델 제출에 대해 설명합니다. 뉴스 번역. 우리는 213 과 같은 기술을 사용했습니다.

대규모 역번역, 바이텍스트 마이닝, 대규모 고밀도 및 희소 다국어 모델, 도메인 미세 조정, 양상블 및 시끄러운 채널 순위 재지정. 우리는 각 기술의 영향을 정량화하기 위해 광범위한 실험 결과를 제공 하고 최종 시스템을 생성하기 위해 누적되는 정도를 제공합니다. 우리의 결과는 다국어 번역이 자원이 적은 언어와 자원이 많은 언어 모두에서 최첨단 성능을 달성할 수 있음을 보여 주며 강력한 이중 언어 기준 과 이전 연도의 수상작을 능가합니다.

감사의 말

Michael Auli와 Halil Akin에게 감사드립니다.

이 프로젝트를 시작하는 데 도움을 주고 그 과정에서 도움과 조언을 해주셔서 감사합니다. 훈련 데이터를 얻는 데 도움을 주신 Holger Schwenk와 Vishrav Chaudhary에게 감사드립니다.

참조

Roei Aharoni, Melvin Johnson 및 Orhan Firat. 2019.

대규모 다국어 신경 기계 번역.

전산 언어학 협회 북미 지부의 2019 회의 진행: 인간 언어 기술, 1 권(긴 논문 및 짧은 논문), 페이지 3874–3884.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry 등 2019. 야생의 대규모 다국어 신경 기계 번역: 결과 및 과제. arXiv 프리프린트 arXiv:1907.05019.

Mikel Artetxe와 Holger Schwenk. 2019. 다국어 문장 임베딩을 사용한 마진 기반 병렬 말뭉치 마이닝. 전산 언어학 협회 제57차 연례 회의 절차, 3197–3203페이지.

로익 바로, 막달레나 비시알스카, 온드레이 보자르, 마르타 R. 코스타-쥬사, 크리스티안 페더만, 이베트 그레이엄, 로만 그룬드키에비츠, 배리 하도우, 마티아스 허크, 에릭 조아니스, 톰 코크미, 필립 코엔, 치키우 로, 니콜라 류베시치, 크리스토프 몬츠, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, Marcos Zampieri. 2020. 기계 번역 (WMT20) 에 관한 2020년 회의 결과. 제5차 기계 번역 회의 절차, 1–55페이지, 온라인. 전산 Linguistics 협회.

Shruti Bhosale, Kyra Yee, Sergey Edunov 및 Michael Auli. 2020. 사전 교육용이 아닌 언어 모델: 빠른 온라인 신경 노이즈 채널 모델링. arXiv 프리프린트 arXiv:2011.07164.

니콜라이 보고이체프와 리코 센리히. 2019. 신경망 기계 번역을 위한 합성 데이터에서 메인, 번역 및 노이즈를 수행합니다. arXiv 프리프린트 arXiv:1911.03362.

Peng-Jen Chen, Jiajun Shen, Matt Le, Vishrav Chaudhary, Ahmed El-Kishky, Guillaume Wenzek, Myle Ott, Marc'Aurelio Ranzato. 2019. Facebook ai의 wat19 미만 마-영어 번역 과제 제출. arXiv 프리프린트 arXiv:1910.06848.

Alexis Conneau, Kartikay Khandelwal, 나만 고알, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, Veselin Stoyanov. 2020. 대규모 비지도 교차 언어 표현 학습. 전산 언어학 협회 제58차 연례 회의 절차, 8440-8451페이지.

Raj Dabre, Atsushi Fujita, Chenhui Chu. 2019. 자원이 적은 신경망 기계 번역을 위한 단계적 미세 조정을 통해 다국어를 활용합니다. 자연어 처리의 경험적 방법에 관한 2019년 회의 및 자연어 처리에 관한 제9회 국제 합동 회의(EMNLP-IJCNLP), 1410-1416페이지.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier. 2018. 대규모 역번역 이해. arXiv 프리프린트 arXiv:1808.09381.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato 및 Michael Auli. 2019. 역번역으로 훈련된 기계 번역 시스템의 평가. arXiv 프리프린트 arXiv:1908.05204.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, Philipp Koehn. 2020. CCAligned: 교차 언어 웹 문서 쌍의 대규모 컬렉션입니다. 프로세스에서 EMNLP의.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary 등. 2021. 영어 중심의 다국어 기계 번역을 넘어. 기계 학습 연구 저널, 22(107):1-48.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan 및 Wei Wang. 2020. 언어에 구애받지 않는 버트 문장 임베딩. arXiv 프리프린트 arXiv:2007.01852.

Orhan Firat, 조경현, Yoshua Bengio. 2016. 공유 주의 메커니즘을 사용한 다자간, 다국어 신경망 기계 번역. 2016년 전산 언어학 협회 북미 지부 회의 진행: 인간 언어 기술, 페이지 866-875.

Yvette Graham, Barry Haddow, Philipp Koehn. 2020. 기계 번역 평가에서의 Translationese. 214 에 관한 2020 회의 절차에서

EMNLP(자연어 처리의 경험적 방법), 72-81페이지.

케네스 히필드. 2011. Kenlm: 더 빠르고 더 작은 언어 모델 쿼리. 통계적 기계 번역에 관한 여섯 번째 워크숍 진행, 187-197페이지.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari 및 Trevor Cohn. 2018. 신경망 기계 번역을 위한 반복 역번역. 신경망 기계 번역 및 생성에 관한 2차 워크숍 진행, 18-24페이지.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomasz Mikolov. 2017. 효율적인 텍스트 분류를 위한 요령 모음. 전산 언어학 협회 유럽 지부의 15차 회의 진행: 2권, 짧은 논문, 427-431 페이지. 전산 언어학 협회.

쿠도 타쿠와 존 리차드슨. 2018. Sentencepiece: 신경 텍스트 처리를 위한 간단하고 언어 독립적인 하위 단어 tokenizer 및 detokenizer. 자연어 처리의 경험적 방법에 관한 2018 회의 진행: 시스템 데모, 66-71페이지.

Dmitry Lepikhin, HyukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, Zhifeng Chen. 2020. Gshard: 조건부 계산 및 자동 샤딩을 사용하여 거대한 모델을 확장합니다. 학습 표현에 관한 국제 회의에서.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer 및 Veselin Stoyanov. 2019. Roberta: 견고하게 최적화된 bert pretraining approach입니다. arXiv 프리프린트 arXiv:1907.11692.

벤자민 마리, 라파엘 루비노, 후지타 아츠시. 2020. 태그 역번역 재검토: 실제로 작동하는 이유는 무엇입니까? 전산 언어학 협회 제58차 연례 회의 절차, 5990-5997페이지.

로버트 무어와 윌리엄 루이스. 2010. 언어 모델 학습 데이터의 지능적인 선택. ACL 2010 Conference Short Papers 진행, 220-224 페이지.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, Sergey Edunov. 2019. [Facebook FAIR의 WMT19 뉴스 번역 작업 제출](#). 기계 번역에 관한 제4차 회의 진행(2권: 작업 공유 문서, 1일차).

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, Michael Auli. 2019. fairseq: 시퀀스 모델링을 위한 빠르고 확장 가능한 툴킷. NAACL-HLT 2019 진행: 데모.

매트 포스트. 2018. [BLEU 점수 보고의 명확성을 요구합니다](#). CoRR, abs/1804.08771.

번역(1권: 연구 논문), 73-81페이지.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, Angela Fan. 2019. CCMatrix: 웹에서 수십억 개의 고품질 병렬 문장 마이닝. arXiv 프리프린트 arXiv:1911.04944.

Micha Ziernski, Marcin Junczys-Dowmunt, Bruno Pouliquen. 2016. 유엔 병렬 코퍼스 v1.0. 언어 자원 및 평가에 관한 제10차 국제 회의(LREC'16) 절차, 3530-3534페이지.

Rico Sennrich, Barry Haddow, Alexandra Birch. 2015. 단일 언어 데이터로 신경망 기계 번역 모델 개선. 전산 언어 학 협회 제54차 연례 회의 절차, 86-96페이지.

Xu Tan, Jiale Chen, Di He, Yingce Xia, QIN Tao, Tie-Yan Liu. 2019. 언어 클러스터링을 사용한 다국어 신경 기계 번역. 자연어 처리의 경험적 방법에 관한 2019년 회의 및 자연어 처리에 관한 제9차 국제 합동 회의(EMNLP-IJCNLP), 962-972페이지.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Na man Goyal, Vishrav Chaudhary, Jiatao Gu 및 Angela Fan. 2020. 확장 가능한 다국어 사전 교육 및 미세 조정을 통한 다국어 번역. arXiv 프리프린트 arXiv:2008.00401.

Jörg Tiedemann. 2012. OPUS의 병렬 데이터, 도구 및 인터페이스. 언어 자원 및 평가에 관한 제8차 국제 회의(LREC'12) 절차, 2214-2218페이지.

Chau Tran, Yuqing Tang, Xian Li, Jiatao Gu. 2020. 반복적인 자기 감독 교육을 위한 교차 언어 검색. 신경 정보 처리 시스템의 발전, 33:2207-2219.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, Illia Polosukhin. 2017. 주의가 필요합니다. 신경 정보 처리 시스템의 발전, 5998-6008페이지.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, Edouard Grave. 2019. Ccnet: 웹 크롤링 데이터에서 고품질 단일 언어 데이터 세트 추출. arXiv 프리프린트 arXiv:1911.00359.

Kyra Yee, Yann Dauphin, Michael Auli. 2019. 신경망 기계 번역을 위한 간단하고 효과적인 노이즈 채널 모델링. 자연어 처리의 경험적 방법에 관한 회의에서.

Biao Zhang, Philip Williams, Ivan Titov, Rico Sennrich. 2020. 대규모 다국어 신경 기계 번역 및 제로샷 번역 개선. 전산 언어학 협회 제58차 연례 회의 절차, 1628-1639페이지.

마이크 장과 안토니오 토랄. 2019. 기계 번역 테스트 세트에서 번역어의 효과. 기계 215 에 대한 네 번째 회의 절차에서