기계번역에서 효율적인 도메인 적응을 위한 적응형 데이터스토어 리사이징

장진희[1], 최주환[2], 윤정민[2], *김영빈[3]
[1]중앙대학교
[2]중앙대학교 AI학과
[3]중앙대학교 첨단영상대학원
oao3398@gmail.com, gold5230@cau.ac.kr,
cocoro357@cau.ac.kr, ybkim85@cau.ac.kr

Efficient Domain Adaptation via Adaptive Datastore Resizing for Machine Translation

Jinhee Jang[1], Juhwan Choi[2], Jungmin Yun[2], and *Youngbin Kim[3]
[1]Chung - Ang University
[2]Department of Artificial Intelligence, Chung - Ang University
[3]Graduate School of Advanced Imaging Science, Multimedia & Film, Chung - Ang University

Abstract

Domain adaptation, a fundamental task in deep learning, aims to mitigate performance degradation caused by domain shifts. In the field of neural machine translation, domain adaptation has been a significant focus, with ongoing research dedicated to addressing this challenge. Existing kNN - MT - based domain adaptation methodologies have been inefficient due to the massive size of the datastore. Subsequent studies attempted to reduce the size of the datastore to shorten search time, but there was a trade - off between size and performance.

In this paper, we propose ARK - MT (Adaptively Resizing Datastore for k - Nearest Neighbor Machine Translation), a methodology that adaptively adjusts the size of the datastore according to the translation model's confidence scores. Experimental findings demonstrate that, compared to existing methods, the ARK - MT model exhibited superior performance across all four domains examined. Moreover, the results validate that ARK - MT enables a reduction in inference time through efficient computational operations.

I. 서론

언어 모델의 발전에 따라 도메인 적응(domain adaptation)의 중요성이 대두되고 있다[1]. 도메인 적응은 특정 도메인에서 학습된 모델을 다른 도메인에 적용할 때 도메인 변환(domain shift)으로 인한 성능 저하를 완화하는 것을 목표로 한다.

기계 번역 분야에서도 이러한 도메인 적응에 대한 연구가 지속적으로 이루어지고 있다[2,3,4,5,6]. 대표적인 연구 중 하나로 최근접 이웃 기계 번역(kNN - MT)[2]은 최근접 이웃(kNN) 검색 알고리즘을 사용하여 추가적인 학습이 필요 없는 도메인 적응 기법을 제안하였다. kNN - MT는 신경망 기계 번역(NMT) 모델이 추론 과정에서 외부 지식을 참고할 수 있도록 해당 도메인에 대한 참조 코퍼스로

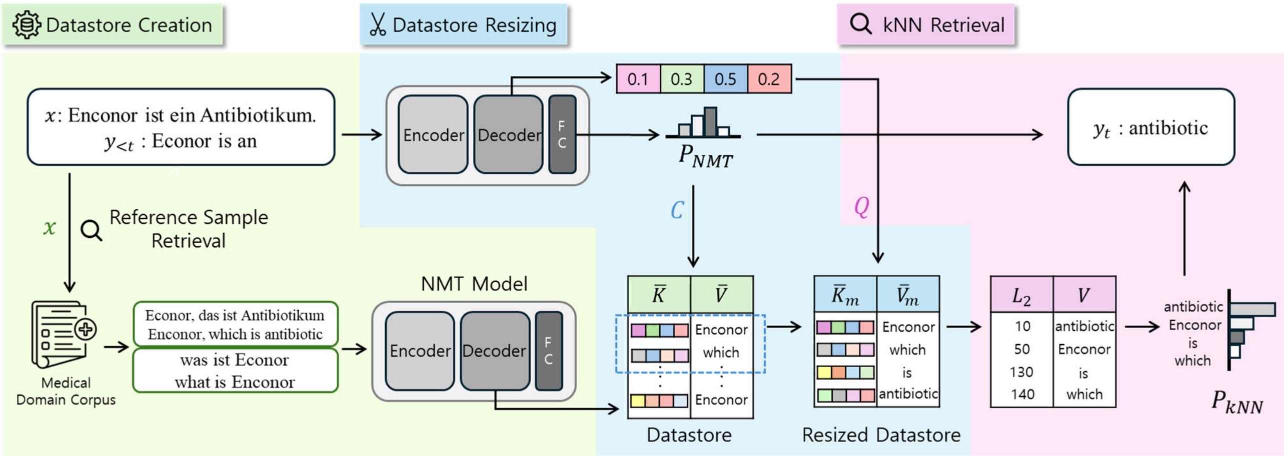

그림 1. ARK-MT의 구조

데이터스토어를 구축하고, 디코딩 시 토큰별 검색을 통해 다음 토큰의 확률 분포를 수정한다. 이를 통해 해당 도메인의 특성을 가진 토큰에 높은 확률을 부여하여 특정 도메인에 특화된 문장의 번역 성능을 크게 개선하였다.

하지만 추가적인 학습을 필요로 하지 않는다는 이점과 성능적 잠재력에도 불구하고, kNN-MT는 거대한 크기의 데이터스토어에 따른 저장 공간의 오버헤드와 검색 연산의 비효율을 야기하여 실제 적용이 어렵다는 한계가 있다.

이러한 단점을 개선하기 위해 kNN-MT의 디코딩 방식을 고도화하여 성능을 개선할 뿐만 아니라 데이터스토어에 대한 검색을 효율화하기 위한 연구들이 이루어졌다[3,5,6]. 특히 데이터스토어의 크기를 줄여 검색 시간을 단축하려는 시도가 주로 이루어졌다.

대표적인 연구인 FK-MT[3]와 SK-MT[5]는 소스(source) 문장을 query로 사용하여 도메인 참조 코퍼스에서 입력 문장과 형태적, 의미적으로 유사한 샘플을 검색한다. 이를 통해 전체 코퍼스 대신 연관성 높은 일부 샘플만을 활용하여 작은 크기의 데이터스토어를 구축함으로써 효율적인 검색이 가능해진다.

하지만 이러한 방식은 정해진 개수의 참조 샘플을 사용하기 때문에 데이터스토어의 크기와 검색 성능 사이의 trade-off가 존재한다. 데이터스토어의 크기를 줄이면 검색의 효율이 향상되지만, 참조 샘플의 수를 줄여 성능 하락을 야기할 수 있다. 이와 반대로 성능 향상을 위해 데이터스토어의 크기를 늘릴 경우 오히려 검색 시간이 증가하는 문제가 발생할 수 있다.

본 논문에서는 고정된 크기의 데이터스토어로 인한 비효율적 연산을 개선하기 위해 적응적 데이터스토어 리사이징(Adaptive Datastore Resizing) 기법을 적용한 ARK-MT (Adaptively Resizing Datastore for k-Nearest Neighbor Machine Translation)을 제안한다. ARK-MT는 토큰 생성 시 번역 모델의 신뢰도(confidence)를 고려하여 데이터스토어의 크기를 적응적으로 조절한다. 이를 통해 검색 속도를 개선하고, 신뢰도가 낮을 경우 더 많은 외부 지식을 고려할 수 있도록 하여 성능 저하를 방지할 수 있다. 또한 모델의 신뢰도에 따라 데이터스토어에 대한 검색 결과를 반영하는 정도를 조정하여 디코딩 시 도메인에 적합한 토큰을 추론할 수 있다.

ARK-MT의 효율성을 검증하기 위해, 일반적인 도메인으로 학습된 번역 모델에 대하여 IT, 의학, 코란, 법률 4가지 도메인을 대상으로 실험을 수행하였다. 실험 결과, 기존 방법론 대비 추론 속도가 증가하면서도 성능이 향상되어 제안한 방법론이 도메인 적응에 효과적임을 확인할 수 있었다.

II. 본론

본 장에서는 기계 번역의 대표적인 도메인 적응 방법론인 kNN-MT를 설명하고 이를 개선하여 도메인 적응의 효율성을 향상시키기 위한 ARK-MT를 소개한다. 제안하는 모델의 구조는 그림 1과 같다.

2.1 데이터스토어 구축

kNN-MT는 최근접 이웃(kNN) 알고리즘을 기반으로 모델이 다음 토큰 생성 시 데이터스토어에 저장된 참조 샘플을 참고하여 트랜스포머 모델의 디코더 출력을 보완한다. 번역을 위한 입력을 x, 출력을 y 라고 하며 특정 도메인의 참조 코퍼스를 $(x, y) \in (X, Y)$ 라고 할 때, 번역 모델 f_θ 은 각 토큰별

생성 시점 t 에 대하여 은닉 상태 $f_\theta(x, y_{<t})$ 를 출력한다. 이는 데이터스토어의 key K 에 해당하며, $(t+1)$의 토큰을 value V로 사용한다. 매 생성 시점의 은닉 상태를 기반으로 데이터스토어 (K, V) 를 구축하며, 이는 아래의 식으로 표현할 수 있다.

$$(K, V) = \bigcup_{(x,y)\in(X,Y)} \{f_\theta(x, y_{<t}), \forall y_t \in y\} \qquad (1)$$

이와 같이 전체 참조 코퍼스를 사용하여 데이터스토어를 구축할 경우 많은 저장공간을 필요로 한다. 예를 들어, 법률 도메인을 구성하는 46만 7천 개의 전체 문장 쌍으로 데이터스토어를 구축할 경우 37GB의 저장공간이 요구되지만, 추론 시 실제로 사용되는 문장은 평균 14.2개에 불과하다[3]. 이는 전체 코퍼스로 데이터스토어를 구축하는 것은 비효율적이며, 참조 코퍼스의 모든 정보가 추론 과정에 기여하지는 않는다는 것을 나타낸다.

이러한 분석을 바탕으로 SK-MT는 BM25[7]과 편집 거리(edit distance)[8]와 같은 간단한 검색 알고리즘을 활용하여 매 문장 생성 시 입력 문장과 관련된 m개의 참조 샘플을 추출하고 이를 기반으로 데이터스토어를 구축한다. SK-MT의 경우 토큰 생성 과정에서 고정된 크기의 데이터스토어를 사용하기 때문에 m이 클수록 더 많은 정보를 참조할 수 있지만 검색 시간이 증가하며, m이 작을수록 검색 시간은 감소하는 반면 성능 하락을 야기할 수 있다.

2.2 kNN 검색

추론 시점 $\hat{t}$까지 생성된 토큰을 $\hat{y}_{<\hat{t}}$라고 할 때, f_θ의 출력인 $f_\theta(x, \hat{y}_{<\hat{t}})$ 는 현재 시점까지의 문맥을 포함한 벡터를 의미한다. 이는 구축한 데이터스토어를 대상으로 kNN 검색을 수행하기 위한 query Q 로 사용된다. 검색 결과 $P_{kNN}(y_{\hat{t}}|x, \hat{y}_{<\hat{t}})$ 는 아래와 같은 식에 따라 검색 분포로 변환된다[3].

$$P_{kNN}(y_{\hat{t}}|x, \hat{y}_{<\hat{t}}) \propto \sum_{(h_i,v_i)} I_{y_t=v_i} \exp\left(\frac{-d\left(h_i, f_\theta(x, \hat{y}_{<\hat{t}})\right)}{\tau}\right) \qquad (2)$$

이때 d 는 유클리디안 거리를 의미하며, h 는 데이터스토어의 key를 나타낸다. τ 는 하이퍼파라미터로, 소프트맥스 함수를 적용한 이후의 확률 분포를 완만하게 조정하여 생성 과정에서 다양한 토큰을 고려할 수 있게 한다.

최종 확률분포인 $P(y_{\hat{t}}|x, \hat{y}_{<\hat{t}})$ 는 검색 확률 분포 P_{kNN} 과 번역 모델의 출력 확률 분포 P_{NMT} 를 모두 반영하며, 이때 하이퍼파라미터 λ를 통해 반영 비율을 결정한다. 해당 식은 다음과 같이 표현할 수 있다.

$$P(y_{\hat{t}}|x, \hat{y}_{<\hat{t}}) = \lambda P_{kNN}(y_{\hat{t}}|x, \hat{y}_{<\hat{t}}) + (1-\lambda) P_{NMT}(y_{\hat{t}}|x, \hat{y}_{<\hat{t}}) \qquad (3)$$

2.3 제안 기법

본 논문에서는 선행 연구의 추론 시간과 성능 간 trade-off를 개선하기 위해 토큰별 모델의 신뢰도에 따라 참조 샘플의 개수 m을 기반으로 데이터스토어의 크기를 적응적으로 조절하는 ARK-MT를 제안한다.

모델의 신뢰도(confidence) c는 번역 모델의 토큰 생성 시점 t 에서의 예측 확률 $P_{NMT}(y_t|x, y_{<t})$ 를 의미하며, 토큰 생성 시 모델의 예측에 대한 확신 수준을 나타내는 지표로 활용된다[9]. 이러한 신뢰도 지표를 기반으로 데이터스토어의 크기를 적응적으로 조절함으로써, c에 따라 모델에 제공하는 외부 정보의 양을 조정할 수 있다. 이를 통해 매 토큰 생성 시 고정된 크기의 데이터스토어를 활용하는 기존 SK-MT 대비 효율적인 검색 연산이 가능하다.

ARK-MT는 도메인 병렬 코퍼스를 대상으로 BM25를 사용하여 연관성이 높은 참조 샘플을 검색한다. 이때, SK-MT와 동일하게 상위 64개의 샘플을 검색 결과로 사용한다. 그 후, 형태적 유사성을 고려하기 위해 편집 거리를 기준으로 검색 결과를 재정렬한다. 검색 결과인 참조 샘플 $(\bar{X}, \bar{Y})$를 기반으로 데이터스토어 $(\bar{K}, \bar{V})$ 를 구축하며, 이는 다음과 같이 표현할 수 있다.

$$(\bar{K}, \bar{V}) = \bigcup_{(\bar{x},\bar{y})\in(\bar{X},\bar{Y})} \{f_\theta(\bar{x}, \bar{y}_{<t}), \forall \bar{y}_t \in \bar{y}\} \qquad (4)$$

데이터스토어는 샘플 단위로 구축되며, 각 토큰 단위의 예측에 대한 c 를 반영하여 m 을 기반으로 데이터스토어의 크기를 적응적으로 조절한다. 이때 참조 샘플의 개수 m은 다음과 같이 정의한다.

$$m = m_{min} + (m_{max} - m_{min})(1-c) \qquad (5)$$

이때 m_{min}과 m_{max}는 m이 될 수 있는 최대 크기와 최소 크기를 뜻한다. 식(5)는 0과 1 사이의 확률값을 m_{max}와 m_{min} 사이로 매핑하며, 이를 통해 낮은 c에 대해서는 높은 m을 도출하고 높은 c에 대해서는 낮은 m 을 도출할 수 있다. 데이터스토어의 크기를 적응적으로 조절하기 위해 참조 샘플 $(\bar{K}, \bar{V})$에 대하여 편집 거리를 기준으로 상위 m 개의 참조 샘플 $(\bar{X}_m, \bar{Y}_m)$ 을 추출한다. 이를 기반으로 데이터스토어 $(\bar{K}_m, \bar{V}_m)$ 를 구축하여 kNN의 수행 범위를 제한하며, 이에 대한 수식은 다음과 같다.

$$(\bar{K}_m, \bar{V}_m) = \bigcup_{(\bar{x}_m,\bar{y}_m)\in(\bar{X}_m,\bar{Y}_m)} \{f_\theta\left(\bar{x}_m, \bar{y}_{m<t}\right), \forall \bar{y}_{m_t} \in \bar{y}_m\} \qquad (6)$$

c가 높아 외부 지식이 상대적으로 불필요한 경우 데이터스토어의 크기를 줄이고, c 가 낮을 경우

데이터스토어의 크기를 늘려 더 많은 외부 지식을 활용한다. 이처럼 토큰별 kNN 검색에서 c 를 고려함으로써 검색 연산량을 효율적으로 조정할 수 있다.

또한, 본 논문에서는 c 를 반영하여 확률 분포의 반영 비율인 λ 의 값을 설정한다. 기존 SK-MT는 kNN 검색 이후, λ값을 통해 최종 확률 분포에 반영될 검색 분포의 비율을 결정한다. 이때 $\lambda = ReLU(1-\frac{d_0}{\tau})$를 통해 Q와 K 간의 최근접 거리를 나타내는 d_0가 클수록 검색 분포의 반영 비율이 낮아지도록 구성하였다.

본 논문에서는 c 를 고려할 수 있도록 λ 를 다음과 같은 식으로 정의한다.

$$\lambda = ReLU(\left(1-\frac{d_0}{\tau}\right)+\alpha(1-c)) \tag{7}$$

이때 α 는 c 를 반영하는 강도를 나타내는 하이퍼파라미터이다. 이를 통해 c가 낮을 경우 검색 분포를 더 많이 반영하여 번역 모델의 출력 확률 분포 P_{NMT}를 보정할 수 있다.

III. 실험

이 장에서는 제안하는 ARK-MT의 효과를 검증하기 위한 실험과 그 결과에 대해 설명한다.

3.1 실험 환경

실험을 위해 WMT 독일어-영어를 우승한 트랜스포머 기반 번역 모델[10]을 사용한다. 데이터셋으로는 다중 도메인 병렬 코퍼스[11]을 사용하였으며 IT, 의료, 코란, 법률 총 4개의 도메인에 대한 번역문으로 구성되어 있다. 이때 2,000쌍의 테스트 데이터셋을 사용하여 실험을 진행하였다.

기존 연구[3]와의 공정한 비교를 위해, m_{min} 과 m_{max} 는 각각 2와 32로 설정하여 m 의 값을 결정하였다. 검색 분포의 반영 정도를 제어하는 두 가중치 파라미터 α와 τ의 값은 각각 0.01과 100으로 설정하였다.

번역 품질에 대한 비교를 위한 평가 지표로는 SacreBLEU[12]를 활용한다. 이를 통해 n-gram 정확도뿐만 아니라 번역문의 길이 적정성과 참조문 대비 중요 단어나 구문 포함 여부 등을 종합적으로 고려할 수 있다.

3.2 실험 결과

ARK-MT를 각 도메인별 데이터셋에 적용하여 도메인 적응을 수행하였으며, 그 결과는 표 1과 같다. kNN 방식을 적용하지 않은 NMT 모델은 kNN-MT 모델에 비해 SacreBLEU 점수가 7.45점이 낮은 성능 격차를 보였다. 한편 제안한 방법론인 ARK-MT는 모든 도메인에 대하여 가장 높은 성능을 보였다. 특히 ARK-MT는 m이 32인 SK-MT 대비 더 적은 수의 참조 샘플을 사용함에도 불구하고 SacreBLEU 점수가 평균 1.98점이 향상된 성능을 보여 신뢰도 기반 제안 방법론의 효과를 확인할 수 있었다.

방법론	m	도메인				평균
		IT	의학	코란	법률	
NMT	-	36.46	41.43	13.97	45.50	34.34
kNN-MT	-	38.06	52.15	15.00	58.73	40.98
SK-MT	8	39.70	52.52	**16.15**	59.11	41.87
	16	**40.43**	**52.69**	15.49	**59.43**	42.01
	32	38.04	51.64	14.70	58.20	40.64
ARK-MT (ours)	2-32	**41.41**	**52.81**	**16.49**	**59.80**	**42.62**

표 1. 도메인별 SacreBLEU 점수 비교

SK-MT는 도메인별로 m 의 크기에 따른 성능 차이가 발생한다. 이에 따라 데이터셋에 대한 최적의 하이퍼파라미터 m 에 대한 탐색이 요구된다. 반면 ARK-MT의 경우 특정 범위 내에서 m 의 값을 적응적으로 조정함으로써 하이퍼파라미터 최적화에 대한 의존도를 상대적으로 낮출 수 있다는 장점이 있다.

또한, SK-MT 대비 ARK-MT의 추론 시간 차이를 비교하였으며 그 결과는 표 2와 같다. 이때 시간 단축 효과를 검증하기 위해 SK-MT의 m은 32, ARK-MT의 m_{max} 는 32로 설정하였다. 실험 결과, 모든 도메인에서 ARK-MT의 추론 시간이 단축되었으며, 추론 속도가 최대 1.27배 향상되었다.

방법론	IT	의학	코란	법률	평균
SK-MT	416	682	864	1052	752
ARK-MT (ours)	**392**	**538**	**682**	**976**	**646**

표 2. SK-MT와 ARK-MT의 추론 시간(초) 비교

SK-MT의 경우 고정된 크기의 데이터스토어를 사용하므로 입력의 개수에 대하여 추론 시간이 선형적으로 증가한다. 반면 ARK-MT는 모델의 신뢰도를 기반으로 외부 지식의 필요 여부를 판별하고 이에 따라 검색 범위를 조정함으로써 효율적인 검색이 가능하다. 이처럼 본 논문에서 제안하는 ARK-MT는 적응적 데이터스토어 리사이징 기법을 기반으로 추론

시간을 단축하면서 성능을 향상시킴을 보여주었다.

IV. 결론 및 향후 연구 방향

본 논문에서는 신경망 기계 번역에서 효율적인 도메인 적응을 위해 모델의 신뢰도를 기반으로 데이터스토어의 크기를 적응적으로 조절하는 ARK-MT를 제안하였다. 실험 결과, 제안 모델은 기존 방법 대비 추론 시간이 단축되면서도 성능이 향상되는 결과를 확인할 수 있었다. 향후 연구에서는 모델의 신뢰도를 참조 샘플의 개수에 매핑하는 방식을 고도화하고, 참조 샘플 검색 시 관련 없는 데이터 샘플을 제거하는 방식을 설계하고자 한다.

Acknowledgements

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(NRF-2022R1C1C1008534)과 정보통신기획평가원의 지원(2021-0-01341, 인공지능대학원지원(중앙대학교))의 지원을 받아 수행된 연구임.

참고문헌

[1] Xu Guo, Han Yu, "On the domain adaptation and generalization of pretrained language models: A survey", arXiv preprint arXiv:2211.03154, 2022.

[2] Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, Mike Lewis, "Nearest neighbor machine translation", In ICLR, 2021.

[3] Yuhan Dai, Zhirui Zhang, Qiuzhi Liu, Qu Cui, Weihua Li, Yichao Du, Tong Xu, "Simple and scalable nearest neighbor machine translation", In ICLR, 2023.

[4] Xin Zheng, Zhirui Zhang, Junliang Guo, Shujian Huang, Boxing Chen, Weihua Luo, Jiajun Chen, "Adaptive nearest neighbor machine translation", In ACL, 2021.

[5] Yuxian Meng, Xiaoya Li, Xiayu Zheng, Fei Wu, Xiaofei Sun, Tianwei Zhang, Jiwei Li, "Fast nearest neighbor machine translation", In ACL (Findings), 2022.

[6] Pedro Henrique Martins, Zita Marinho, André F. T. Martins, "Efficient Machine Translation Domain Adaptation", In Proc. ACL 2022 Workshop on Semiparametric Methods in NLP: Decoupling Logic from Knowledge. 2022.

[7] Robertson, Stephen, Hugo Zaragoza, "The probabilistic relevance framework: BM25 and beyond", In Foundations and Trends® in Information Retrieval 3.4 : 333-389. 2009.

[8] Navarro, Gonzalo, "A guided tour to approximate string matching", In ACM computing surveys (CSUR) 33.1 : 31-88. 2001.

[9] Charles Corbiere, Nicolas THOME, Avner Bar-Hen, ` Matthieu Cord, Patrick Perez, "Addressing failure prediction by learning model confidence", In Advances in Neural Information Processing Systems, 2019.

[10] Philipp Koehn, Rebecca Knowles, "Six challenges for neural machine translation", In First Workshop on Neural Machine Translation (WNMT), 2017.

[11] Nicolas Stroppa, Antal van den Bosch, Andy Way, "Exploiting source similarity for smt using context-informed features", In Proceedings of The 11th Conference on Theoretical and Methodological Issues in Machine Translation, 2007.

[12] Matt Post, "A call for clarity in reporting bleu scores", ArXiv, abs/1804.08771, 2018.