**KIPS Trans. Softw. and Data Eng.**
Vol.12, No.8 pp.341~354
ISSN: 2287-5905 (Print), ISSN: 2734-0503 (Online)


https://doi.org/10.3745/KTSDE.2023.12.8.341

# Cross-Lingual Style-Based Title Generation Using Multiple Adapters

Yo-Han Park† · Yong-Seok Choi†† · Kong Joo Lee†††

## ABSTRACT

The title of a document is the brief summarization of the document. Readers can easily understand a document if we provide them with its title in their preferred styles and the languages. In this research, we propose a cross-lingual and style-based title generation model using multiple adapters. To train the model, we need a parallel corpus in several languages with different styles. It is quite difficult to construct this kind of parallel corpus; however, a monolingual title generation corpus of the same style can be built easily. Therefore, we apply a zero-shot strategy to generate a title in a different language and with a different style for an input document. A baseline model is Transformer consisting of an encoder and a decoder, pre-trained by several languages. The model is then equipped with multiple adapters for translation, languages, and styles. After the model learns a translation task from parallel corpus, it learns a title generation task from monolingual title generation corpus. When training the model with a task, we only activate an adapter that corresponds to the task. When generating a cross-lingual and style-based title, we only activate adapters that correspond to a target language and a target style. An experimental result shows that our proposed model is only as good as a pipeline model that first translates into a target language and then generates a title. There have been significant changes in natural language generation due to the emergence of large-scale language models. However, research to improve the performance of natural language generation using limited resources and limited data needs to continue. In this regard, this study seeks to explore the significance of such research.

Keywords : Title Generation, Cross-lingual, Cross-style, Multiple Adapter

# 다중 어댑터를 이용한 교차 언어 및 스타일 기반의 제목 생성

박 요 한† · 최 용 석†† · 이 공 주†††

## 요 약

문서의 제목은 문서의 내용을 가장 효율적으로 요약하여 제공해 준다. 이때 독자들이 선호하는 스타일과 언어에 따라 문서의 제목을 다르게 제공해 준다면, 독자들은 문서의 내용을 좀 더 쉽게 예측할 수 있다. 본 연구에서는 문서가 주어졌을 때 언어와 스타일에 따라 제목을 자동 생성하는 '교차 언어 및 스타일 기반의 제목 생성 모델'을 제안한다. 모델을 학습하기 위해서는 같은 내용을 다른 언어와 다른 스타일로 작성한 병렬데이터가 필요하다. 그러나 이러한 종류의 병렬데이터는 구축하기 매우 어렵다. 반면, 단일 언어와 단일 스타일로 구축된 제목 생성 데이터는 많으므로 본 연구에서는 제로샷(zero-shot) 학습으로 제목 생성을 수행하고자 한다. 교차 언어 및 스타일 기반의 제목 생성을 학습하기 위해 다중 언어로 사전 학습된 트랜스포머 모델에 각 언어, 스타일, 기계번역을 위한 어댑터를 추가하였다. 기계 번역용 병렬데이터를 이용하여 기계번역을 먼저 학습한 후, 동일 스타일의 제목 생성을 학습하였다. 이때, 필요한 어댑터만을 학습하고 다른 부분의 파라미터는 모두 고정시킨다. 교차 언어 및 스타일 기반의 제목을 생성할 때에는 목적 언어와 목적 스타일에 해당하는 어댑터만을 활성화시킨다. 실험 결과로는 각 모델을 따로 학습시켜 파이프라인으로 연결시킨 베이스라인에 비해 본 연구에서 제안한 제로샷 제목 생성의 성능이 크게 떨어지지 않았다. 최근 대규모 언어 모델의 등장으로 인한 자연어 생성에서의 많은 변화가 있다. 그러나 제한된 자원과 제한된 데이터만을 이용하여 자연어 생성의 성능을 개선하는 연구는 계속되어야 하며, 그런 점에서 본 연구의 의의를 모색한다.

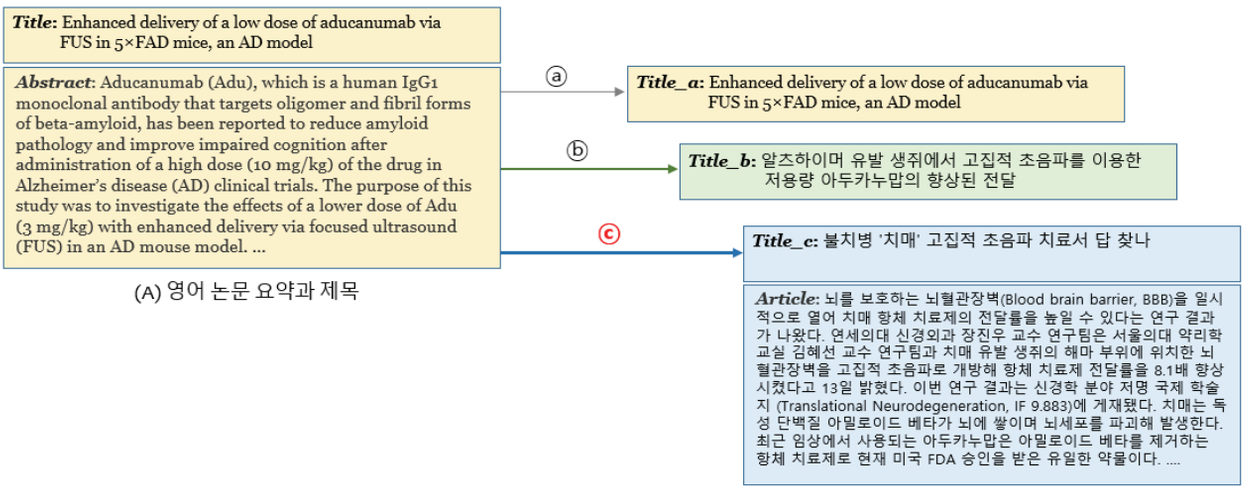
키워드 : 제목 생성, 교차 언어, 교차 스타일, 멀티 어댑터

## 1. 서 론

문서의 내용을 효율적으로 제공할 수 있는 대표적인 방법은 내용을 요약한 제목을 제공하는 것이다. 독자는 제목을 통해 문서를 읽기 전에 내용의 대략적인 정보를 예상할 수 있다.

논문과 신문기사는 제목을 통해 내용을 최대한 요약하여 제공하는 문서들이다. 논문은 특정 주제나 이론 등을 전문가

※ 본 논문은 2022년도 교육부의 재원으로 한국연구재단의 지원을 받아 수행된 지자체-대학 협력기반 지역혁신 사업의 결과임(2021RIS-004).
† 비 회 원 : 충남대학교 전파정보통신공학과 석 · 박사통합과정
†† 준 회 원 : 충남대학교 전자전파정보통신공학과 박사과정
††† 종신회원 : 충남대학교 전파정보통신공학과 교수
Manuscript Received : February 16, 2023
First Revision : April 24, 2023
Accepted : May 18, 2023
* Corresponding Author : Kong Joo Lee(kjoolee@cnu.ac.kr)




(A) 영어 논문 요약과 제목

(B) 한국어 신문 기사와 제목

Fig. 1. Example of Title Generation

Table 1. Notation of Title Generation Task

| TASKS | sub-TASKS | notation |
|---|---|---|
| ⓐ Title Generation | Title Generation (TG) | $Doc \rightarrow Title$ |
| ⓑ Cross-lingual Title Generation | Title Generation (TG) + Machine Translation (MT) | $Doc^{En} \rightarrow Title^{Ko}$<br>$Doc^{Ko} \rightarrow Title^{En}$ |
| **ⓒ Cross-lingual Style-based Title Generation** | Title Generation (TG) + Machine Translation (MT) + Style Transfer (ST) | $Doc^{En}_{paper} \rightarrow Title^{Ko}_{News}$<br>$Doc^{Ko}_{News} \rightarrow Title^{En}_{paper}$<br>$Doc^{En}_{News} \rightarrow Title^{Ko}_{Paper}$<br>$Doc^{Ko}_{paper} \rightarrow Title^{En}_{News}$ |

가 논리적으로 작성한 글로, 제목에 전문 용어와 축약어가 사용된다. 해당 분야의 전문가인 경우에는 논문의 제목으로 대략의 내용을 예상할 수 있지만, 일반 독자들의 경우에는 제목만으로 내용을 예측하는 것은 매우 어려운 일이다. 또한, 대부분의 학술 논문은 영어로 작성된 경우가 많아 비영어권 독자들에게는 접근성이 떨어진다. 반면 신문기사는 기자가 사실을 바탕으로 작성한 글로, 제목은 간결하면서도 독자들이 쉽게 이해할 수 있도록 작성된다. 또한 모국어로 작성된 경우가 많아 일반 독자가 쉽게 접할 수 있다. 이와 같이 독자들의 관련 지식 정도와 사용 언어에 따라 문서의 제목을 다르게 제공하면, 독자들은 문서의 내용을 보다 쉽게 예측할 수 있다.

따라서 본 연구에서는 문서가 주어졌을 때 독자가 선호하는 언어와 제목 스타일에 따라 제목을 자동 생성하는 `교차 언어 및 스타일 기반의 제목 생성 모델'을 제안한다. 이때 언어는 한국어와 영어를 대상으로 하며, 제목 스타일은 논문과 신문기사의 제목 스타일을 대상으로 한다. 예를 들어, 영어 학술 논문에 대해 신문기사 스타일로 작성된 한국어 제목을 제공하는 것이다.

교차 언어로 스타일에 따른 제목 생성 모델을 학습하기 위해서는 여러 언어로 작성된 논문과 신문기사의 병렬 데이터가 필요하다. 하지만 동일한 내용을 다루고 있는 논문과 신문기사의 병렬 데이터는 매우 드물며, 교차 언어로 구성된 데이터는 더더욱 구하기 어렵다. 그러나 논문-제목 데이터와 신문기사-제목 데이터는 매우 쉽게 구할 수 있다. 본 연구에서는 교차 언어 및 스타일 기반의 제목을 생성하기 위해 영어-논문-제목으로 구성된 데이터와 한국어-신문기사-제목으로 구성된 데이터만을 이용하여 원하는 언어와 스타일의 제목을 제로샷(zero-shot) 학습으로 생성해 본다.

Fig. 1[1]의 (A)는 영어로 작성된 의학 분야 논문의 초록과 제목이며 (B)는 (A)논문의 내용을 소개한 한국어 신문 기사와 제목이다. 본 연구의 목적은 문서의 내용을 요약하여 제목을 생성하는 것이다. 이때, 작업 ⓐ는 원문과 동일한 언어로 제목을 생성한 것이고, 작업 ⓑ는 원문과 다른 언어로 제목을 생성한 것이며, 작업 ⓒ는 교차 언어이면서 다른 스타일로 제목을 생성한 것이다. Table 1은 Fig. 1의 제목 생성 작업과 본 논문에서 사용할 표기 방법을 정리한 것이다. 본 연구는

1) 논문 출처: https://translationalneurodegeneration.biomedcentral.com/articles/10.1186/s40035-022-00333-x
신문기사 출처: https://www.medicaltimes.com/Main/News/NewsView.html?ID=1151591

작업 ©의 교차 언어 및 스타일 기반의 제목 생성을 목적으로 한다. 교차 언어로는 한국어와 영어를 적용하며, 제목 스타일은 학술 논문과 신문 기사 두 종류의 스타일을 적용한다.

작업 ©의 Title_c는 신문기사의 제목 스타일로 전문용어를 최대한 배제하고 독자가 흥미를 느낄 수 있도록 작성된 것이다. 작업 ©를 교차 언어와 스타일의 모든 조합에 대하여 학습시키기 위해서는 [영어-논문-제목, 한국어-신문기사-제목]으로 구성된 병렬 데이터뿐만 아니라 [한국어-논문-제목, 영어-신문기사-제목]의 병렬 데이터도 존재해야 한다. 하지만 이런 병렬 데이터는 매우 구하기 어렵기 때문에 본 연구에서는 제로샷 학습 방법을 도입하여 제목 생성을 수행한다.

우리가 본 연구에서 사용할 수 있는 데이터는 (1) 영어-논문-제목 데이터, (2) 한국어-신문기사-제목 데이터, (3) 영어-한국어 병렬 데이터이다. 이 세 종류의 데이터만을 이용하여 제로샷의 교차 언어 및 스타일 기반의 제목을 생성한다. 그렇기 때문에 본 연구에서는 영어 논문을 입력으로 한국어 신문기사 스타일의 제목 생성($Doc^{En}_{paper} \rightarrow Title^{Ko}_{News}$)과 한국어 신문기사를 입력으로 영어 논문 스타일의 제목 생성($Doc^{Ko}_{News} \rightarrow Title^{En}_{paper}$)만을 다루고, $Doc^{En}_{News} \rightarrow Title^{Ko}_{Paper}$과 $Doc^{Ko}_{paper} \rightarrow Title^{En}_{News}$의 제목 생성은 배제한다.

[1]에서는 데이터가 적은 언어에 대해 제로샷 또는 퓨샷(few-shot)을 수행할 때 고자원 언어의 정보를 효율적으로 사용하기 위한 프레임워크인 MAD-X를 제안하였다. MAD-X는 언어에 따른 어댑터와 작업(task)를 위한 어댑터를 구분하여 각 언어와 작업에 특화된 모듈을 구축하고자 하였다. 본 연구에서는 MAD-X에 아이디어를 얻어 다국어 사전 훈련 인코더-디코더 모델인 mBART[2]에 언어 어댑터와 스타일 어댑터를 추가하여 제로샷 제목 생성을 수행하였다.

## 2. 관련 연구

### 2.1 Adapter

BERT[3] 개발 이후 대량의 데이터로 모델을 사전 학습(Pre-training)한 뒤 소량의 데이터로 미세 조정(Fine-tuning)하는 것이 자연어 처리의 표준 관행이 되었다. 하지만 작업 종류가 많아질 경우 모든 작업에 대해 사전 학습 모델을 따로 미세 조정하는 것은 비효율적이다. 이를 해결하기 위해 [4]에서는 각 작업을 학습할 소량의 매개 변수를 모델에 추가하는 어댑터(adapter)를 제안하였다.

어댑터는 down-projection와 up-projection의 2층의 Multi-layer Perceptron으로 구성되며 트랜스포머의 모든 레이어에 추가된다. 어댑터는 전체 사전 학습 모델에 비해 훨씬 적은 매개변수로 구성되어 있다. 또한 각 작업별로 어댑터만을 학습시키고 사전 학습 모델은 학습시키지 않기 때문에 모든 작업에 대해 사전 학습 모델의 매개변수를 공유할 수 있다.

### 2.2 MAD-X

다국어 사전 학습 모델(Multilingual Pretrained Model)은 한 모델에 여러 언어를 동시에 사전 학습시킨다. 다국어 사전 학습 모델은 학습데이터가 많은 고자원 언어와 데이터 수가 적은 저자원 언어를 함께 학습함으로써 저자원 언어의 성능 향상을 목적으로 한다. 그러나 모델 파라미터 수의 제한으로 인해 저자원 언어 또는 사전 학습에 포함되지 않은 언어의 성능은 여전히 낮다. [1]은 이러한 문제를 해결하기 위해 언어별 어댑터와 작업별 어댑터를 함께 사용하는 Muliple ADapter for Cross-lingual transfer(MAD-X)를 제안하였다.

MAD-X는 다국어 사전 학습 인코더 모델인 mBERT의 각 레이어에 어댑터를 추가하여 케추아어에 대한 제로샷 개체명 인식(Named Entity Recognition, NER)을 수행하였다. 언어 어댑터로는 학습데이터가 많은 영어 어댑터와 목표 언어인 케추아어의 어댑터를 추가하였으며, 작업 어댑터로는 목표 작업인 개체명인식 어댑터를 사용하였다.

모델 학습은 두 단계로 이루어진다. 첫 번째 단계에서는 영어와 케추아어의 텍스트 데이터를 이용하여 Masked Language Model을 학습한다. 영어 데이터에 대해서는 영어 어댑터만 활성화하고, 케추아어 데이터는 케추아어 어댑터만 활성화하여 학습한다. 이때, mBERT의 매개변수는 학습하지 않는다. 두 번째 단계에는 영어 개체명 인식 데이터만을 학습하는데, 영어 어댑터와 케추아어 어댑터, mBERT의 매개변수는 고정한 채로, 개체명인식 어댑터만을 학습한다. 학습 이후 케추아어에 대한 개체명 인식을 수행할 때에는 케추아어 어댑터와 개체명인식 어댑터를 활성화하여 최종 예측을 수행한다.

[1]에서는 저자원 언어의 단어 임베딩 학습이 미흡하기 때문에 고자원 언어와 저자원 언어에 대한 단어 정렬을 위한 Invertible Adapter를 추가하였다. Invertible adapter는 트랜스포머의 제일 아래 층에 위치하고 트랜스포머의 제일 상위 층에는 Inversed adapter를 사용한다. 이때 Invertible adapter와 Inversed adapter는 사전 학습 모델의 정보를 읽지 않기 위해 서로 역함수 관계가 되어야 한다. 이를 만족하기 위해서 [1]에서는 Non-linear Independent Component Estimation (NICE)를 적용하였다.

### 2.3 제로샷을 통한 교차 제목 생성

독자들은 제목을 통해 문서 전체의 대략적인 정보를 얻을 수 있기 때문에 독자가 선호하는 언어나 스타일로 제목을 제공해 주는 것이 중요하다. 하지만 동일한 문서에 대한 교차 속성의 제목 데이터를 수집하는 것은 매우 어려우므로 대부분의 교차 제목 생성은 제로샷 방법의 연구가 진행되고 있다.

[5]의 연구에서는 영어 문서를 입력으로 중국어 헤드라인을 생성하는 Cross-lingual Neural Headline Generation (CNHG)를 제안하였다. 제로샷 학습을 위해 영어-중국어 기계 번역 모델을 구축하고 이를 선생 모델로 사용한다. 영어 헤

드라인을 중국어로 기계 번역하여 정답으로 사용하였다. 영어 문서를 CNHG에 입력하여 중국어 헤드라인을 생성하고, 기계 번역으로 나온 중국어 헤드라인과의 KL-divergence를 통해 모델을 학습하였다.

[6]의 연구에서는 동일한 문서에 대한 서로 다른 스타일의 제목을 생성하는 Stylistic Headline Generation(SHG)를 제안하였다. SHG는 헤드라인 생성과 오토인코더를 동시에 학습하는 멀티 태스크를 통해 학습하였다. 이때 오토인코더는 각 스타일의 데이터를 사용하였으며, 모델이 스타일을 구분하여 텍스트를 생성하기 위해 Style-Guided Encoder Attention과 Style-Dependent Layer Norm을 디코더에 추가하여 사용하였다.

## 3. 교차 언어 및 스타일 기반의 제로샷 제목 생성

### 3.1 모델

본 연구에서 제안하는 교차 언어 및 스타일 기반의 제로샷 제목 생성 모델은 Fig. 2와 같다. 제목 생성 모델은 인코더-디코더로 구성되며, 문서를 입력으로 받아 그 내용을 요약한 제목을 생성한다. 인코더-디코더 모델은 다국어 사전학습 모델인 mBART[4]를 베이스모델로 사용하고 인코더와 디코더에 어댑터를 추가하였다. 어댑터는 크게 언어 어댑터와 스타일 어댑터로 나눌 수 있다. 언어 어댑터로는 영어(Lang:En) 어댑터와 한국어(Lang:Ko) 어댑터를 사용한다. 스타일 어댑터는 본 연구에서 생성하고자 하는 논문 제목(Style:Paper)과 신문기사 제목(Style:News)의 두 가지 어댑터를 사용한다.

어댑터는 [3]과 같이 down-projection과 up-projection의 두 개의 층으로 이루어져 있으며, Equation (1)은 어댑터 층의 수식이다.

$$Adapter(h,r) = U(ReLU(D(h))) + r \tag{1}$$

$D$와 $U$는 각각 down과 up-projection 층이며, $h$는 어댑터 입력 벡터이다. 어댑터의 residual connection($r$)은 기존 mBART의 정보가 손실되지 않게 트랜스포머 피드 포워드(feed forward)의 출력을 사용하였다.

모델을 학습시키기 위해서는 [영어-논문-제목, 한국어-신문기사-제목]의 병렬 데이터가 필요하지만 이러한 데이터를 수집하는 것은 매우 어렵다. 그러므로 본 연구에서는 영어-논문-제목 데이터와 한국어-신문기사-제목 데이터를 따로 활용하여 제목 생성 모델을 학습한다. 하지만 이렇게 두 종류의 데이터로만 학습할 경우 영어로는 논문 스타일의 제목만, 한국어로는 신문기사 스타일의 제목만을 생성하는 단점이 있다. 본 연구에서는 교차 언어 정보를 학습할 수 있도록 기계 번역 학습에 사용하는 기계 번역 어댑터(MT)를 추가하였다.

[1]에서와 같이 다국어 사전 학습 모델의 임베딩 정보를 위해 인버터블 어댑터도 사용하였다. 인버터블 어댑터는 트랜스포머 최하단에 invertible과 최상단에 inversed 두 개를 함께 사용한다. Equation (2)는 invertible 어댑터의 수식이며, $e_1$와 $e_2$는 각 토큰의 워드 임베딩($e$)을 같은 크기를 갖도록 나눈 벡터이다.

$$\begin{aligned} &A_{\in v}(e) = o = [o_1, o_2] \\ &o_1 = F(e_2) + e_1; o_2 = G(o_1) + e_2 \end{aligned} \tag{2}$$

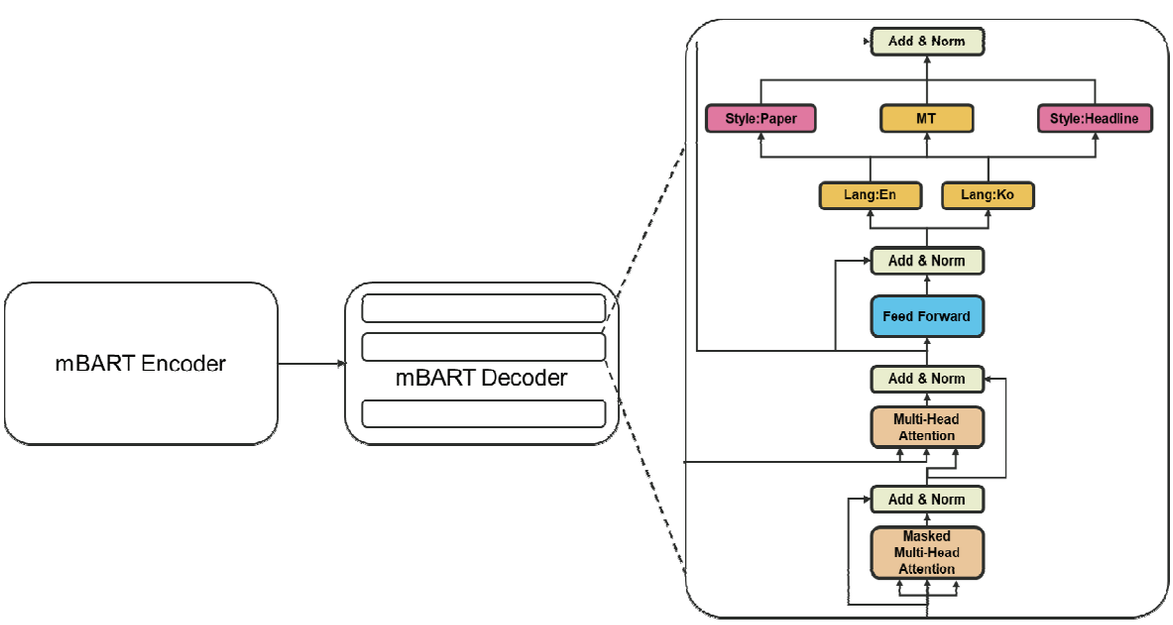



Fig. 2. Cross-lingual Style-based Title Generation Model

Inversed 어댑터는 invertible 어댑터의 출력을 복원하는 역할을 한다. Equation (3)은 inversed 어댑터의 수식이다.

$$A_{\in v}^{-1}(o) = e = [e_1, e_2]$$
$$e_2 = o_2 - G(o_1); e_1 = o_1 - F(e_2) \quad (3)$$

인코더에는 인버터블 어댑터와 언어 어댑터만 추가되며, 디코더에는 인버터블 어댑터, 언어 어댑터, 기계 번역 어댑터, 스타일 어댑터가 포함된다.

### 3.2 학습 방법

교차 언어 및 스타일 기반의 제목 생성 모델의 학습은 두 단계로 수행된다. 제목 생성 모델을 학습하는 과정은 Fig. 3과 같다.

**Step 1:**

첫 번째 단계에서는 기계 번역 작업을 수행하여 두 언어 사이의 정렬을 학습한다. mBART는 다국어 텍스트 데이터로 사전학습이 되었지만, 병렬 데이터가 아닌 단일 언어 코퍼스로 학습되었기 때문에 두 언어 사이의 정렬이 잘 이루어지지 않았다. 한국어-영어 병렬 데이터를 이용하여 인코더-디코더의 언어 어댑터와 기계 번역 어댑터를 같이 학습하며, 인버터블 어댑터도 학습한다. 한국어→영어 번역을 학습할 때에는 인코더의 한국어 어댑터, 디코더의 기계 번역 어댑터와 영어 어댑터를 활성화 시키고, 반대로 영어→한국어 번역을 수행할 때에는 인코더의 영어 어댑터, 디코더의 기계 번역 어댑터와 한국어 어댑터를 활성화하여 학습한다.

**Step 2:**

두 번째 단계에서는 영어-논문-제목 데이터와 한국어-신문기사-제목 데이터를 사용하여 스타일 기반의 제목 생성을 학습한다. 이때 첫 번째 단계에서 학습한 언어 정보가 소실되지 않도록 인코더-디코더의 언어 어댑터의 파라미터는 업데이트하지 않고 디코더의 제목 스타일 어댑터만을 학습한다. 이 단계에서는 기계 번역 어댑터는 사용하지 않는다.

학습 후 추론과정에서 영어-논문을 한국어-신문기사-제목으로 생성하는 방법은 Fig. 4와 같다. 그림에서 보는 바와 같이 영어-논문이 입력으로 들어오면 인코더의 영어 어댑터를 활성화하고, 디코더의 한국어 어댑터와 신문기사 제목 스타일 어댑터를 활성화시켜 한국어 신문기사 스타일의 제목을 제로샷으로 생성한다.

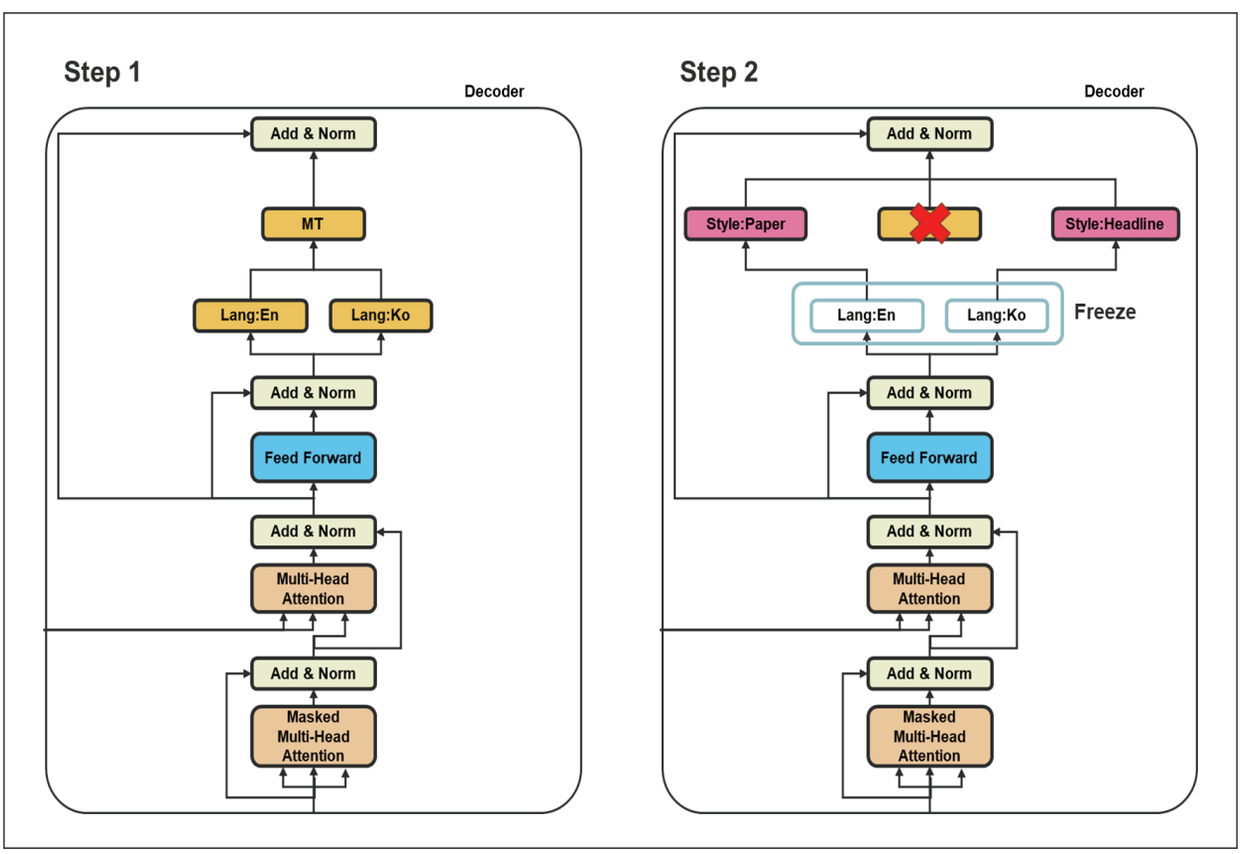



Fig. 3. Training Process of Cross-lingual Style-based Title Generation Model

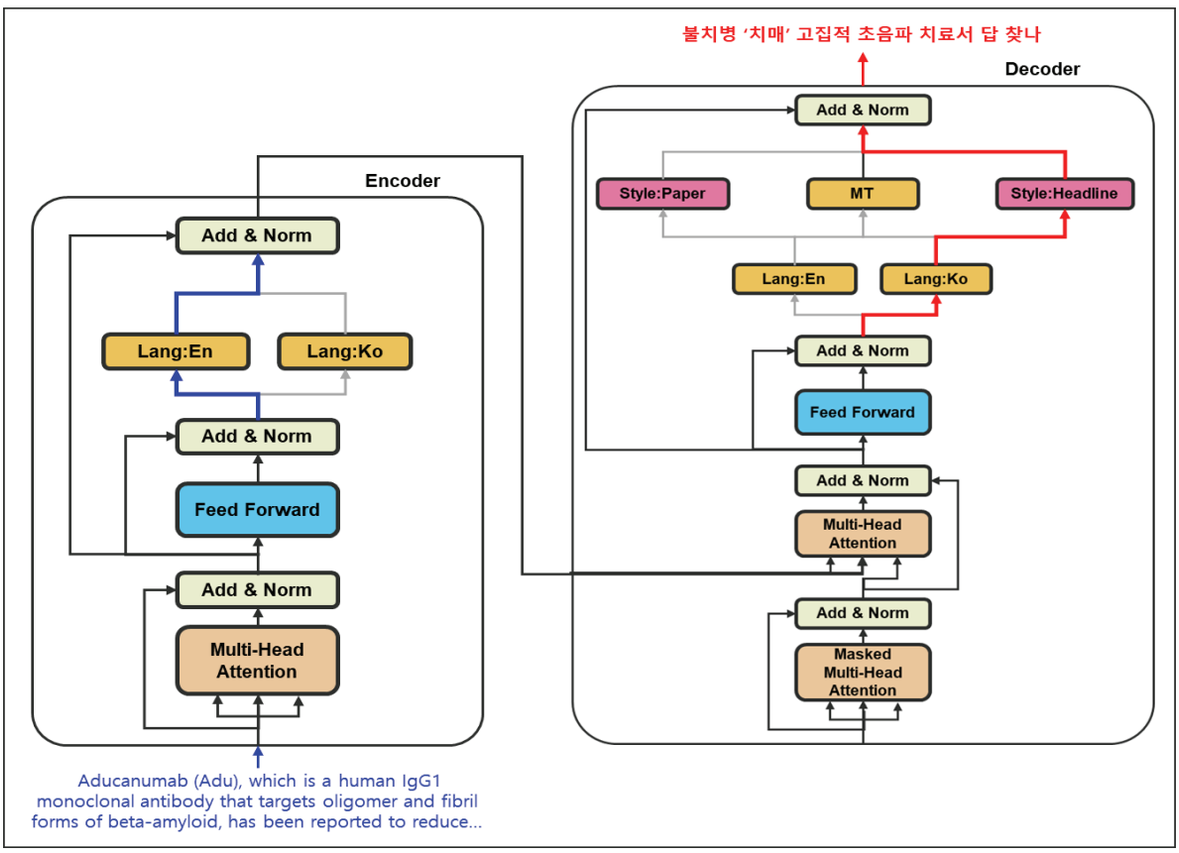



Fig. 4. Cross-lingual Style-based Title Generation(English Paper Document → Korean News Headline)

## 4. 실 험

### 4.1 데이터 집합

본 연구에서는 실험을 위해 의학 계열의 영어 학술 논문과 한국어 의학 분야 신문 기사를 수집하였다.

영어 논문은 PubMed(https://pubmed.ncbi.nlm.nih.gov/)에서 제공한 2021 baseline 버전으로부터 임의의 204,656개의 문서를 선택하여 데이터를 수집하였다. PubMed는 미국 국립생물공학정보센터(National Center for Biotechnology Information; NCBI)에서 1997년 6월부터 무료로 제공하고 있는 데이터베이스로 생의학 전반의 최신 의학 서지 정보를 검색할 수 있는 의학 분야 최고의 학술 검색 사이트이다. 모든 서지 정보는 고유의 PMID로 식별이 가능하다.

한국어 신문 기사는 의학 뉴스 사이트[2]로부터 2020년 10월 23일까지 게시된 신문 기사 221,298개를 수집했다.

검증과 평가를 위해서 동일한 내용을 담고 있는 [영어-논문-제목, 한국어-신문기사-제목]의 병렬 데이터가 필요하다. 이를 위해 생명과학 전공자 1명, 약학 전공자 2명, 총 3명의 대학생이 2020년 4월부터 8주간 총 447시간에 걸쳐 수동으로 2,658개의 병렬 데이터를 직접 구축하였으며 구축 과정은 다음과 같다.

첫째 한국어 의학 분야 뉴스사이트로부터 새로운 연구 결과를 소개하는 신문기사를 찾는다. 신문기사 내용 중, 병원 소개나 학회 개최 소개 등 학술 논문 내용과 무관한 것은 모두 제거한다. 둘째, 신문기사 내용 중, 원래 논문의 출처 정보가 있는 것만을 남기고 나머지는 모두 제거한다. 셋째, 저자와 저널 정보를 이용하여 원래 논문을 찾고 해당 논문의 PMID값을 추출한다.

Fig. 5는 데이터 구축과정의 예시이다. Fig. 5A는 한국어 신문기사 페이지이며 기사 말미에 학술지명이 제시되어 있다. 학술지 명과 저자 이름을 키워드로 Fig. 5B와 같은 영어 원 논문을 검색하고 해당 논문의 PMID를 추출한다. 신문기사에는 학술지 이름만 명시되어 있고 영어 원 논문의 제목이 명확히 제시되어 있지 않은 경우가 많아 해당 분야를 전공하는 학생들이 영어 원제목을 유추하여 검색하는 과정이 필요하였다. Appendix A에 수집된 데이터의 일부를 제시하였다.

2) www.natureasia.com
www.khanews.com
www.ibric.org
www.monews.co.kr
www.sciencetimes.co.kr
www.medicaltimes.com
www.cancerline.co.kr
www.bosa.co.kr news.healthi.kr
www.medical-tribune.co.kr

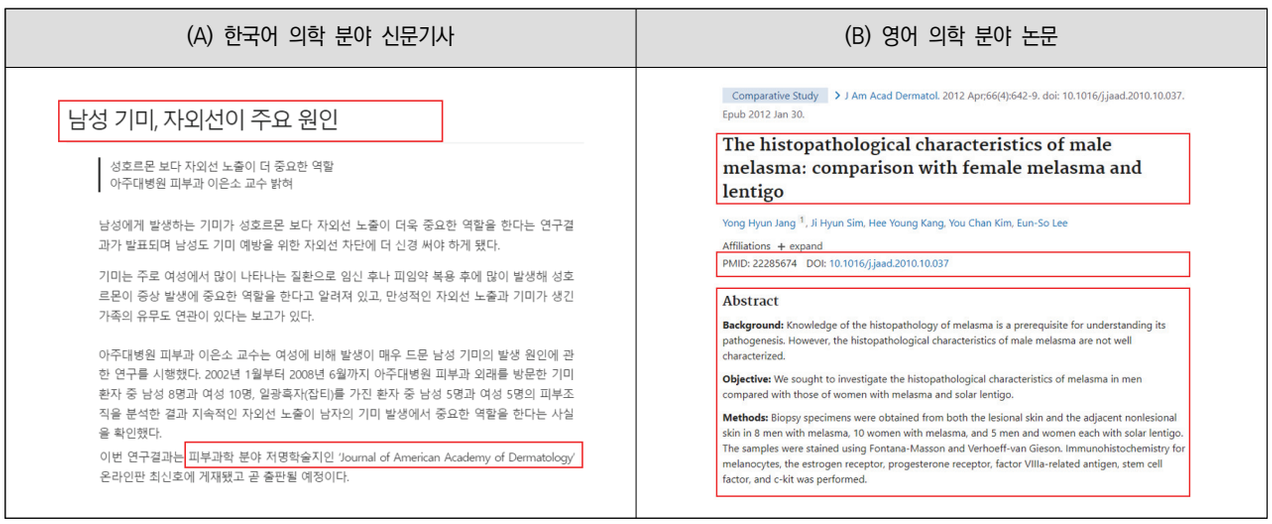



Fig. 5. Construction Process of Evaluation Dataset

Table 2. Example of Augmented Translation Data

| Augment Type | # | Lang | Medical Vocabulary | Augmented Data |
|---|---|---|---|---|
| AEDA | (1) | Ko | 힘줄주위염 | ! , ? ? ; **힘줄주위염** ! ; ; , !" |
| | | En | peritendinitis | ! , ? ? ; **peritendinitis** ! ; ; , ! |
| | (2) | Ko | 췌장머리암 | ? , . , ? ! , : ! **췌장머리암** ; |
| | | En | pancreas head cancer | ? , . , ? ! , : ! **pancreas head cancer** ; |
| Sentence | (3) | Ko | 췌장통증 | 이것은 **췌장통증** 입니다. |
| | | En | pancreatalgia | It is **pancreatalgia**. |
| | (4) | Ko | 대동맥 주위 림프절 | 그것은 **대동맥 주위 림프절** 입니다. |
| | | En | paraaortic lymph node | That is **paraaortic lymph node**. |

본 연구에서 사용한 제목 생성 데이터는 의학 계열의 데이터이다. 첫 번째 학습 단계인 기계 번역을 학습할 때, 학습 데이터의 도메인을 일치시키기 위해 AIHUB[3]에서 제공하는 '전문분야 한국어-영어 말뭉치' 중 의료/보건 분야와 '한국어-영어 번역 말뭉치(기술과학)'에서 의학 분야의 병렬 데이터를 수집하였으며, 총 673,843개의 병렬데이터를 기계 번역 학습데이터로 사용하였다.

제한된 양의 병렬 데이터만으로는 의학 용어의 번역을 충분히 학습할 수 없다. 그러므로 본 연구에서는 대한의사협회[4]와 서울아산병원의 의학 용어 사이트[5]로부터 114,143개의 의학 용어에 대한 번역쌍을 수집하여 기계 번역과 함께 학습하였다. 하지만 기계 번역은 문장 단위의 번역이기 때문에 단어 단위의 번역과는 불일치가 생긴다. 단어 수준의 번역을 문장 단위로 확장하기 위해 두 가지 데이터 증강 기법을 사용하였다. 첫 번째 증강 기법은 [7]에서 제안한 AEDA(An Easier Data Augmentation)이다. AEDA는 '.', ',', '!', '?', ';', ':'의 6개 문장부호를 무작위로 추가하여 문장을 만드는 것이다. 두 번째 증강 기법은 간단한 패턴을 이용하여 단어를 문장화 하는 것이다. 번역 단어쌍을 문장으로 만들기 위해 "이것은 ...입니다."와 "그것은 ...입니다."의 두 가지 패턴을 이용하여 단어를 문장으로 변환하였다. Table 2는 번역어를 문장화하는 예제이다.

학습을 위해 114,143개의 의학 용어 번역쌍에서 두 개의 AEDA 기법과 두 가지 패턴의 문장화기법을 적용하여 총 456,572개의 증강데이터를 구축하였다.

Table 3은 실험에 사용한 기계 번역 및 제목 생성 데이터의 통계이다. 기계 번역 데이터는 문장 수이고, 제목 생성 데이터는 문서 수이다.

### 4.2 실험 환경

본 연구에서 사용한 하이퍼 파라미터는 Table 4와 같다. 트랜스포머의 하이퍼 파라미터는 mBART-large와 동일하다. 어댑터의 down과 up-projection의 히든 레이어의 크기는 트랜스포머 임베딩 크기의 절반인 512를 사용하였다.

3) https://aihub.or.kr/
4) http://term.kma.org/search/list.asp
5) https://www.amc.seoul.kr/asan/healthinfo/easymediterm/easyMediTermList.do
https://www.amc.seoul.kr/asan/healthinfo/symptom/symptomSubmain.do

Table 3. Experimental Data Statistics

| Task | Data type | | Number of Data |
|---|---|---|---|
| Machine Translation | Train | Parallel | 673,843 |
| | | Augment | 456,572 |
| | Valid | | 64,744 |
| | Test | | 49,745 |
| Title Generation | Train | Eng-Paper-Title | 221,298 |
| | | Kor-News-Title | 204,656 |
| | Valid | [Eng-Paper-Title, Kor-News-Title] (parallel) | 723 |
| | Test | | 1,935 |

Table 4. Hyper-parameters for Model and Training

| | Hyper-parameters | Values |
|---|---|---|
| mBART | Embedding dimension | 1,024 |
| | Feed-forward dimension | 4,096 |
| | Encoder / Decoder layers | 12 |
| | Attention heads | 8 |
| | Activation Function | Relu |
| Criterion | Label smoothed cross entropy | 0.2 |
| Learning parameters | Optimizer | Adam |
| | eps | 1e-6 |
| | Betas | (0.9, 0.98) |
| | Learning rate scheduler | polynomial decay |
| | Learning rate | 0.001 |
| | Warmup updates | 3,000 |
| | Total updates | 20,000 |
| | Max tokens | 3,072 |
| | Dropout | 0.3 |
| | Attention dropout | 0.1 |
| | Update Frequency | 8 |

### 4.3 평가 지표

성능 비교를 위해 세 가지 평가 지표를 사용하였다. 첫 번째는 기계 번역에서 주로 사용되는 BLEU Score[8]이다. 정답 제목과 생성된 제목 사이에 1부터 4까지의 n-gram이 얼마나 겹치는지 측정한다. 두 번째 지표는 문서 요약에서 사용되는 ROUGE score[9]이다. ROUGE score는 uni-gram(1), bi-gram(2), longest(L) 세 가지를 측정한다.

BLEU와 ROUGE score는 생성된 제목과 정답 제목 사이에 동일한 단어가 얼마나 사용되었는지로 성능을 평가한다. 그렇기 때문에 정답과 다른 단어로 작성된 비슷한 의미의 제목은 제대로 평가하지 못한다. 이를 보완하기 위하여 본 연구에서는 BERTScore[10]도 평가 지표로 사용하였다. BERTScore는 생성된 문장과 정답 문장을 BERT에 입력하여 각각의 문장 임베딩을 추출하고 두 임베딩의 코사인 유사도를 지표로 두 문장의 의미가 얼마나 유사한지 측정한다.

## 5. 실험 결과

본 연구에서 제안한 모델의 성능 비교를 위해 베이스라인 모델을 다음과 같이 설정하였다.

4.1절의 학습데이터를 이용하여 영-한($MT^{E\rightarrow K}$)과 한-영($MT^{K\rightarrow E}$) 기계 번역기를 각각 구축하고, 영어-논문-제목 데이터를 이용하여 논문 스타일의 제목 생성기($TG_{Paper}$)와 한국어-신문기사-제목 데이터를 이용하여 신문기사 스타일의 제목 생성기($TG_{News}$)을 구축하였다. 이후 각 모델을 파이프라인(pipeline)으로 연결하여 베이스라인 모델로 삼았다. 두 개의 번역기와 두 개의 제목 생성기 모두 mBART를 기반으로 구축되었으며, 해당 작업을 위한 미세조정은 디코더에 어댑터 하나만 추가하여 수행하였다.

### 5.1 제목 생성 모델

Table 5는 영어 논문 또는 한국어 신문기사가 입력으로 주

어졌을 때 제목 생성에 대한 성능 비교이다. 즉, 언어나 스타일의 변화 없이 생성된 제목을 평가한 것이다. 베이스라인 모델은 각 스타일의 제목 생성기를 독립 모델로 구축한 반면, 본 연구에서 제안한 모델은 서로 다른 두 스타일의 제목을 어댑터를 이용하여 하나의 모델에 같이 구축하였다. 그러므로, 제목을 생성할 때 베이스라인 모델은 두 개의 서로 다른 모델이 필요하지만, 본 연구에서 제안한 모델은 하나의 모델에서 활성화하는 어댑터만을 바꿔가며 다른 스타일의 제목을 생성할 수 있다.

또한 제안한 모델은 학습 첫 단계에서 기계 번역 과제를 학습하였다. 실험 결과 영어 논문 제목 생성의 경우, 베이스라인 모델($TG_{Paper}$)이 제안한 모델(Ours)보다 더 좋은 성능을 보였다. 반면 한국어 신문 기사 제목 생성의 경우 본 연구에서 제안한 모델이 베이스라인 모델에 비해 큰 폭의 성능 향상을 보였다. 이는 사전 학습 데이터가 적은 한국어 생성의 경우 첫 단계의 기계 번역 작업이 성능 향상에 도움을 준 것으로 추정된다.

### 5.2 교차 언어 및 스타일 기반의 제목 생성 모델

Table 6은 언어와 스타일을 모두 다르게 하여 생성한 제목에 대한 성능 비교이다. 'w/o augment'는 첫 번째 단계인 기계 번역을 학습할 때 증강데이터를 사용하지 않은 모델이다. 베이스라인 모델인 MT→$TG_{Style}$ 은 원문을 기계 번역한 후, 타겟 스타일의 제목을 생성하는 것이고, TG→MT은 입력 문서와 동일 스타일로 제목을 생성한 후, 이를 기계 번역하여 최종 제목을 생성하는 것이다.

먼저 증강데이터 사용 유무에 따른 성능을 비교하였을 때, 한국어 신문기사 스타일의 제목 생성의 경우 증강데이터를 사용하였을 때 성능이 향상되었다. 반면, 영어 논문 스타일의 제목 생성에서는 증강데이터를 사용하지 않았을 때의 성능이 더 우수했다. 한국어보다 훨씬 더 많은 영어 문서를 사전 학습한 mBART에게 다소 부자연스럽게 만들어진 증강데이터는 오히려 영어 문장 생성에 방해 요소로 작용한 듯하다.

제목 생성 후 번역을 수행(TG→MT)한 결과는 한국어 신문기사 스타일의 제목 생성과 영어 논문 스타일의 제목 생성의 두 경우 모두 제일 낮은 성능을 보였다. 이는 스타일 변환 없이 교차 언어로만 제목을 생성했기 때문이다.

한국어 신문 기사 제목 생성에서는 번역 후 신문기사 스타일의 제목 생성($MT^{E\rightarrow K}\rightarrow TG_{News}$)을 한 것이 제안한 모델(Ours)보다 좋은 성능을 보였다. 그러나 영어 논문 제목에서는 본 연구에서 제안한 모델이 더 좋은 성능을 보였다.

Table 5. Performance Comparison of Title Generation

| Task | Model | BLEU | | | | ROUGE | | | BERT Score |
|---|---|---|---|---|---|---|---|---|---|
| | | B-1 | B-2 | B-3 | B-4 | 1 | 2 | L | |
| $Doc^{En}_{Paper}\rightarrow Title^{En}_{Paper}$ | Ours | 23.10 | 9.10 | 4.40 | 2.30 | 32.50 | 14.42 | 27.62 | **0.93** |
| | TGPaper (baseline) | **24.30** | **9.90** | **5.00** | **2.70** | **35.77** | **17.93** | **30.92** | 0.90 |
| $Doc^{Ko}_{News}\rightarrow Title^{Ko}_{News}$ | Ours | **31.80** | **15.10** | **7.60** | **3.90** | **9.42** | **3.40** | **9.41** | 0.83 |
| | TGNews (baseline) | 20.70 | 7.10 | 2.70 | 1.30 | 6.06 | 1.17 | 5.99 | **0.84** |

Table 6. Performance Comparison of Cross-lingual Style-based Title Generation

| Task | Model | BLEU | | | | ROUGE | | | BERT Score |
|---|---|---|---|---|---|---|---|---|---|
| | | B-1 | B-2 | B-3 | B-4 | 1 | 2 | L | |
| $Doc^{En}_{Paper}\rightarrow Title^{Ko}_{News}$ | Ours | 16.10 | 3.70 | 1.20 | 0.40 | 7.43 | 2.75 | 7.22 | 0.81 |
| | Ours (w/o augment) | 15.80 | 3.90 | 1.30 | 0.40 | 6.96 | 2.50 | 6.88 | 0.82 |
| | TGPaper → MTE→K | 5.20 | 0.60 | 0.10 | 0.00 | 6.23 | 1.41 | 6.03 | **0.85** |
| | MTE→K → TGNews | **17.2** | **4.90** | **1.70** | **0.70** | **9.18** | **4.20** | **9.09** | 0.82 |
| $Doc^{Ko}_{News}\rightarrow Title^{En}_{Paper}$ | Ours | 16.10 | 4.00 | 1.50 | 0.60 | 23.51 | 6.80 | 19.67 | 0.93 |
| | Ours (w/o augment) | **16.90** | **4.50** | **1.80** | **1.00** | **24.21** | **7.26** | **20.20** | 0.92 |
| | TGNews → MTK→E | 11.6 | 2.00 | 0.50 | 0.10 | 13.24 | 2.71 | 11.38 | 0.92 |
| | MTK→E → TGPaper | 15.50 | 4.10 | 1.70 | 1.00 | 22.64 | 6.61 | 19.06 | **0.93** |

전담 작업만을 학습한 모델을 파이프라인으로 연결하여 만든 베이스라인은 보통 제로샷 학습과의 비교에서는 강력한 비교모델로 간주된다. 또한 본 연구에서 제안한 모델은 하나의 모델로 교차 언어 및 스타일이 변환된 제목을 생성할 수 있는 반면, 베이스라인의 경우에는 각각 2개, 모두 4개의 학습 모델이 필요하다.

현재 수집된 데이터는 영어-논문-제목, 한국어-신문기사-제목과 같이 언어와 제목 스타일이 쌍으로 존재한다. 그러다 보니 멀티 어댑터로 구축한 모델에서 다양한 어댑터의 조합에 따른 충분한 학습이 이루어지지 못하였다. 또한, 실질적인 활용의 측면에서 보았을 때, 한국어 신문기사 스타일의 제목 생성만이 유의미한 결과로 보이는 측면이 있다. 그러나 영어-논문-제목, 한국어-신문기사-제목뿐만 아니라 영어-신문기사-제목과 한국어-논문-제목의 데이터도 구축하여 멀티 어댑터를 충분히 학습시킨다면 성능 향상이 가능할 것으로 생각한다. 이는 추후 연구에서 보완할 예정이다.

Table 7은 본 연구에서 제안한 모델의 제목 생성 예제이다. 예제 (1)의 한국어 뉴스 제목에서는 “systemic lupus erythematosus”을 “전신 루푸스”로 적절히 번역하여 생성하였고, 영어 논문 제목에도 핵심 단어인 “systemic lupus erythematosus”을 포함하여 생성하였다. (2)번 예제에서는 “당뇨병에 기여”라는 다소 어색한 표현의 제목을 생성하였는데, 그 의미는 정답 제목인 “당뇨병 발병 증가”와 유사하게 해석될 수 있다.

(3)번 예제에서는 전문 용어인 ‘Oxidized phospholipids (OxPL)’을 “인지질 산화”로 번역하지 못하고 축약어인 “OxPL”을 신문기사 스타일의 제목에 그대로 사용하였다.

Table 7. Example of Cross-lingual Style-based Title Generation

| | | |
|---|---|---|
| (1) | Paper Abstract | Cardiovascular disease is an important complication in patients with **systemic lupus erythematosus (SLE)**. Variant alleles of the mannose-binding lectin gene are associated with SLE as well as with severe atherosclerosis. We determined whether mannose-binding lectin variant alleles were associated with an increased risk of arterial thrombosis among patients with SLE. Mannose-binding lectin alleles were genotyped by means of a polymerase-chain-reaction assay in 91 Danish patients with SLE. Arterial and venous thromboses occurring after the diagnosis of SLE were assessed in a prospective study. Arterial and venous thromboses were confirmed by appropriate diagnostic methods.<br>... |
| | News Article | 덴마크 Bispebjerg대학병원 류마티스내과 Tommy Ohlenschlaeger박사가 실시한 새로운 연구에 의하면 전신성 홍반성낭창(systemic lupus erythematosus, SLE)환자에서 혈액단백질의 일종인 만노스결합렉틴(mannose-binding lectin:MBL)변이 대립유전자(allel)의 호모접합성은 동맥혈전증 위험을 높인다고 New England Journal of Medicine(2004;351:260-267)에 발표됐다. 정맥혈전증은 무관 Ohlenschlaeger박사가 91명의 SLE환자를 전향적으로 평가한 결과 동맥혈전증 발병이 MBL변이 대립유전자의 O/O유전자형 환자 84명 중 18명에서 나타났다. 정맥혈전증은 NBL유전자형과는 통계학적으로 무관했다. 정맥혈전증 위험은 어떠한 대립유전자를 갖고 있어도 증가하지 않으며 MBL이 동맥혈전증을 예방하는데 특이적인 역할을 갖고 있는 것으로 시사됐다.<br>... |
| | Paper Title | Mannose-binding lectin variant alleles and the risk of arterial thrombosis in systemic lupus erythematosus |
| | **Generated Paper Title** | Increased risk of ***systemic lupus erythematosus*** in rheumatoid arthritis patients: a prospective cohort study in the New England Journal of Medicine. |
| | News Title | SLE의 동맥혈전증에 MBL유전자형 관련 |
| | **Generated News Title** | 심혈관질환, ***전신 루푸스*** 색색전증 유전자형과 연관 |
| (2) | Paper Abstract | Immunoglobulin E (IgE) is known to activate mast cells. Prior studies have shown that mast cells contribute to diet-induced obesity and diabetes mellitus (DM). We aimed to determine whether adults with IgE sensitization were at risk of DM. We performed assays regarding serum total IgE and allergen-specific IgE levels against the house dust mite, the cockroach, and the dog on 1,528 adults randomly sampled from every age and gender group in various districts. The total and three allergen-specific IgE levels were positively correlated with fasting glucose level and insulin resistance. Subjects with increased levels of total IgE (>100 kU/L), compared to those without, had an odds ratio (OR) of 1.72 (95% confidence interval [95% CI], 1.17-2.54) for DM after adjusting for various covariates.<br>... |

| | | |
|---|---|---|
| (2) | News Article | 알레르기성 질환의 면역에 관여하는 알레르기면역글로불린 E 항체의 농도가 증가할수록 당뇨병 등 대사증후군 발생이 높아진다는 연구가 국내 최초로 입증됐다. 연구에 의하면 바퀴벌레나 집먼지 진드기에 해당 항체가 반응할 경우 당뇨병 발병 위험이 증가할 수 있으므로 집안을 청결하게 해 당뇨병 발병 위험을 최소화 하는 것이 좋다. 가톨릭대학교 여의도성모병원 내분비내과 권혁상(교신저자)·김미경 교수(제1저자) 연구팀이 국민건강영양조사를 바탕으로 총 1528명(남 755명, 여 733명)을 대상으로 혈중 총 IgE 농도와 주요 흡입 알레르기 유발 항원인 집먼지진드기, 바퀴벌레에 대한 특이 IgE농도를 혈당수치와 비교 분석했다. 대표적 알레르기면역글로불린 E(Immunoglobulin E, 이하 IgE) 알레르기성 질환의 면역에 관여하는 항체로 알려져 있다. ... |
| | Paper Title | House dust mite and Cockroach specific Immunoglobulin E sensitization is associated with diabetes mellitus in the adult Korean population |
| | **Generated Paper Title** | Increased risk of developing metabolic syndrome in patients with high blood glucose and high blood pressure due to allergy |
| | News Title | 알레르기 항체 높을수록, ***당뇨병 발병 증가*** 최초 입증 |
| | **Generated News Title** | 면역글린E, 비만·***당뇨병에 기여*** |
| (3) | Paper Abstract | **Oxidized phospholipids (OxPL)** are ubiquitous, are formed in many inflammatory tissues, including atherosclerotic lesions, and frequently mediate proinflammatory changes. Because OxPL are mostly the products of non-enzymatic lipid peroxidation, mechanisms to specifically neutralize them are unavailable and their roles in vivo are largely unknown. We previously cloned the IgM natural antibody E06, which binds to the phosphocholine headgroup of OxPL, and blocks the uptake of oxidized low-density lipoprotein (OxLDL) by macrophages and inhibits the proinflammatory properties of OxPL. Here, to determine the role of OxPL in vivo in the context of atherogenesis, we generated transgenic mice in the Ldlr-/- background that expressed a single-chain variable fragment of E06 (E06-scFv) using the Apoe promoter. E06-scFv was secreted into the plasma from the liver and macrophages, and achieved sufficient plasma levels to inhibit in vivo macrophage uptake of OxLDL and to prevent OxPL-induced inflammatory signalling. ... |
| | News Article | 산화된 저 밀도 지질단백질(low-density lipoprotein, LDL)은 동맥경화가 유발된 부위에 축적된다. 산화된 인지질은 염증을 유발하는 것으로 알려져 있으며, 이러한 인지질이 동맥경화를 유발한다고 생각되고 있다. Joseph Witztum과 공동연구원들은 동맥경화에 대한 산화 가설을 in vivo에서 입증한 연구 결과를 발표하였다. 저자들은 산화된 인지질을 특이적으로 인지하고 제거하는 항체를 발현하는 유전자 조작 생쥐 모델을 이용하여, 산화된 인지질이 고 콜레스테롤 혈증을 가진 생쥐에서 동맥경화 유발을 돕는다는 사실을 확인하였다. |
| | Paper Title | Oxidized phospholipids are proinflammatory and proatherogenic in hypercholesterolaemic mice |
| | **Generated Paper Title** | Oxidative hypothesis of oxidized hypodensity lipoprotein induces hypercholesterolemia in mice with hypercholesterolemia. |
| | News Title | **인지질 산화**와 동맥경화증 과의 상관관계 확인 |
| | **Generated News Title** | OxPL 저밀도 지질 과산화 억제 |

## 6. 결 론

본 연구에서는 입력 문서를 사용자가 원하는 언어 또는 스타일의 제목으로 생성하는 교차 언어 및 스타일 기반의 제목 생성 모델을 제안하였다. 서로 다른 언어와 스타일이 병렬로 되어 있는 데이터의 수집은 매우 어려우므로 제로샷 기반의 제목 생성 모델을 구축하였다. 다국어 사전 학습 모델인 mBART에 스타일 어댑터와 언어 어댑터를 두 층으로 추가하여 모델을 구축하였다. 이후 모델을 학습할 때에는 기계 번역 데이터를 이용하여 먼저 언어 어댑터를 학습한 후, 제목 생성 모델 데이터를 이용하여 스타일 어댑터를 추가 학습하였다.

제목 생성의 대상 문서가 의학 분야이기 때문에 의학 용어에 대한 번역어 학습을 보강하기 위해 데이터 증강을 수행하였다. 기계 번역은 주로 문장 단위로 수행되기 때문에 의학 용어 번역 쌍을 AEDA와 간단한 문장화 기법을 이용하여 번역 데이터로 증강시켰다.

실험 결과 제안한 모델은 번역 후 스타일별 제목 생성을 수행한 베이스라인보다 영어 논문 제목 생성은 우수한 반면, 한국어 신문기사 제목 생성에서는 다소 성능이 떨어지는 것을 볼 수 있었다.

현재 제안한 모델은 영어는 논문 제목 스타일만, 한국어는 신문기사 제목 스타일만을 생성한다. 차후 데이터 보강을 통해 언어별, 제목 스타일별 모든 조합에 대하여 제목 생성이 가능하도록 확장할 계획이다.

최근 OpenAI에서 출시한 ChatGPT가 자연어 처리를 넘어 인공지능 분야에 새로운 역사를 쓰고 있다. 동시에 자연어 생성 연구 분야의 패러다임 전환이 필요하다. 그러나 제한된 자원과 제한된 데이터만을 이용하여 최대의 품질을 얻어야 하는 자연어 처리의 요구 조건은 여전히 유효하며 그런 점에서 본 연구의 의의를 모색한다.

## References


[1] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "MAD-X: An adapter-based framework for multi-task cross-lingual transfer," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[2] Y. Liu et al., "Multilingual denoising pre-training for neural machine translation," *Transactions of the Association for Computational Linguistics*, Vol.8, pp.726-742. 2020.

[3] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT*, 2019.

[4] N. Houlsby et al., "Parameter-efficient transfer learning for NLP," *International Conference on Machine Learning. PMLR*, 2019.

[5] S. Q. Shen, Y. Chen, C. Yang, Z. Y. Liu, and M. S. Sun, "Zero-shot cross-lingual neural headline generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol.26, No.12, pp.2319-2327, 2018.

[6] D. Jin, Z. Jin, J. T. Zhou, L. Orii, and P. Szolovits, "Hooks in the headline: Learning to generate headlines with controlled styles," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[7] A. Karimi, L. Rossi, and A. Prati, "AEDA: An easier data augmentation technique for text classification," Findings of the Association for Computational Linguistics: EMNLP 2021. 2021.

[8] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLUE: a method for automatic evaluation of machine translation," *Proceedings of the 40th Annual Meeting of the ACL*, pp.311-318, 2002.

[9] C.-Y. Lin, "ROUGE: A package for automatic evalu ation of summaries," *Proceedings of the ACL-04 Workshop*, 8, 2004 .

[10] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," *International Conference on Learning Representations (ICLR)*, 2020.



**박 요 한**

https://orcid.org/0000-0002-5023-5604
e-mail : happy115012@cnu.ac.kr
2020년 충남대학교 전파정보통신공학과 (학사)
2020년 ~ 현 재 충남대학교 전파정보통신공학과 석 · 박사통합과정
관심분야 : 자연언어처리, 기계학습, 인공지능



**최 용 석**

https://orcid.org/0000-0002-7889-8004
e-mail : yseokchoi@cnu.ac.kr
2016년 충남대학교 정보통신공학과(학사)
2016년 ~ 2017년 충남대학교 전자전파정보통신공학과(석사)
2018년 ~ 현 재 충남대학교 전자전파정보통신공학과 박사과정
관심분야 : 자연언어처리, 정보검색, 기계학습, 인공지능



**이 공 주**

https://orcid.org/0000-0003-0025-4230
e-mail : kjoolee@cnu.ac.kr
1992년 서강대학교 전자계산학과(학사)
1994년 한국과학기술원 전산학과(석사)
1998년 학국과학기술원 전산학과(박사)
1998년 ~ 2003년 한국마이크로소프트(유) 연구원
2003년 이화여자대학교 컴퓨터학과 대우전임강사
2004년 경인여자대학 전산정보과 전임강사
2005년 ~ 현 재 충남대학교 전파정보통신공학과 교수
관심분야 : 자연언어처리, 기계학습, 인공지능, 정보검색

## Appendix A. 평가데이터 예제

Table A.1 평가 데이터 예제

| | | |
|---|---|---|
| (1) | Paper Title | **The histopathological characteristics of male melasma : comparison with female melasma and lentigo** |
| | Paper Abstract<br><br>**PMID:** 22285674 | Knowledge of the histopathology of melasma is a prerequisite for understanding its pathogenesis. However, the histopathological characteristics of male melasma are not well characterized. We sought to investigate the histopathological characteristics of melasma in men compared with those of women with melasma and solar lentigo. Biopsy specimens were obtained from both the lesional skin and the adjacent nonlesional skin in 8 men with melasma, 10 women with melasma, and 5 men and women each with solar lentigo. The samples were stained using Fontana-Masson and Verhoeff-van Gieson. Immunohistochemistry for melanocytes, the estrogen receptor, progesterone receptor, factor VIIIa-related antigen, stem cell factor, and c-kit was performed. Increased vascularity was found in the lesion of male melasma. The lesion to nonlesion ratio of the vessel area was increased in male melasma compared with lentigo groups. In the lesion of male melasma, there was a significant increase of stem cell factor and c-kit expression. In addition, the lesion to nonlesion ratio of stem cell factor was increased in male melasma compared with female melasma and lentigo groups. The lesion to nonlesion ratio of c-kit was also increased in male melasma compared with lentigo groups. This study did not include clinical data regarding social habits and was not confirmed by other molecular techniques. The results suggest that chronic ultraviolet radiation associated with signaling of paracrine cytokines plays an important role in the mechanism associated with hyperpigmentation in male melasma. |
| | News Title | **남성 기미, 자외선이 주요 원인** |
| | News Article | 남성에게 발생하는 기미가 성호르몬 보다 자외선 노출이 더욱 중요한 역할을 한다는 연구결과가 발표되며 남성도 기미 예방을 위한 자외선 차단에 더 신경 써야 하게 됐다. 기미는 주로 여성에서 많이 나타나는 질환으로 임신 후나 피임약 복용 후에 많이 발생해 성호르몬이 증상 발생에 중요한 역할을 한다고 알려져 있고, 만성적인 자외선 노출과 기미가 생긴 가족의 유무도 연관이 있다는 보고가 있다. 아주대병원 피부과 이은소 교수는 여성에 비해 발생이 매우 드문 남성 기미의 발생 원인에 관한 연구를 시행했다. 2002년 1월부터 2008년 6월까지 아주대병원 피부과 외래를 방문한 기미 환자 중 남성 8명과 여성 10명, 일광흑자(잡티)를 가진 환자 중 남성 5명과 여성 5명의 피부조직을 분석한 결과 지속적인 자외선 노출이 남자의 기미 발생에서 중요한 역할을 한다는 사실을 확인했다. 남성 기미 환자의 병변 부위에서 만성적인 자외선 노출에 의해 생기는 '일광탄력섬유증'과 '진피 속 혈관의 증식' 정도가 병변에 인접한 정상부위보다 증가하는 경향을 관찰한 것. 특히 만성적인 자외선 노출에 의해 유도되고 피부의 색소침착을 유발할 수 있는 '줄기세포인자'가 병변의 표피와 섬유아세포 주변부에서 진하게 염색됐으며, 그의 수용체인 'c-kit' 역시 병변 부위의 표피층에서 인접 정상부위에 비해 의미 있게 증가하는 결과를 보였다. 또한 이러한 변화의 정도는 다른 질환군(여성 기미, 남성 및 여성 일광흑자)과 비교했을 때에도 남성 기미 환자에서 의미 있게 증가함을 확인했다. 반면 기존 연구에서 여성의 기미와 연관이 있다고 알려진 성호르몬인 에스트로겐, 프로게스테론 수용체의 발현 정도는 남성 기미에서 병변 부위와 정상 부위에서 의미 있는 차이가 없었던 만큼 남자 기미의 증상 발생에서는 자외선 노출이 성호르몬보다 더 중요한 역할을 한다는 것을 볼 수 있었다. 이은소 교수는 "이번 연구결과는 현재까지 구체적 연구가 거의 진행된 적이 없는 남성 기미의 발생에서 자외선 노출이 성호르몬에 비해 주요한 원인임을 밝혔다는 데 의의가 있는 것으로, 난치성 색소질환인 남성 기미의 발생과 악화를 예방하는데 가장 중요한 요소가 자외선 노출 차단이 될 것"이라며 "향후 남성 기미의 발생과 관련해 유전적 요인 등 다른 원인에 대한 추가적인 연구가 필요할 것"이라고 밝혔다. 이번 연구결과는 피부과학 분야 저명학술지인 'Journal of American Academy of Dermatology' 온라인판 최신호에 게재됐고 곧 출판될 예정이다. |
| (2) | Paper Title | **Morning hypertension and night non-dipping in patients with diabetes and chronic kidney disease** |
| | Paper Abstract<br><br>**PMID:** 26311166 | Morning hypertension (HTN) and nocturnal non-dipping (ND) are closely associated with target organ damage and cardiovascular events. However, their importance in diabetics with advanced renal disease is unclear. We evaluated the relationships of morning HTN and ND with estimated glomerular filtration rate (eGFR) and proteinuria, and determined the risk of morning HTN and ND according to presence of diabetes mellitus (DM) and chronic kidney disease (CKD) stage. A total of 1312 patients, including 439 with diabetes, were prospectively recruited at 21 centers in Korea. All patients had HTN and an eGFR of 15-89 ml min(-1) per 1.73 m(2). Ambulatory 24-h blood pressure was assessed. The rates of morning HTN (25.2% vs. 13.6%, P<0.001) and ND (58.2% vs. 48.2%, P=0.002) were higher in diabetics than in non-diabetics. eGFR was correlated with ND in all patients (P<0.05) and with morning HTN only in non-diabetics (P=0.005). Proteinuria was related to ND in all patients (P<0.05) and to morning HTN only in diabetics (P=0.001). In a regression analysis, the risk of morning HTN was 2.093 (95% confidence interval (95% CI): 1.070-4.094) for the DMCKD2 group, 1.634 (95% CI: 1.044-2.557) for the CKD3-4-only group and 2.236 (95% CI: 1.401-3.570) for the DMCKD3-4 group compared with the CKD2-only group. The risk of ND was high for stage 3-4 CKD: 1.581 (95% CI: 1.180-2.120) for non-diabetics and 1.842 (95% CI: 1.348-2.601) for diabetics. Diabetics showed higher rates of morning HTN, ND and uncontrolled sustained HTN compared with non-diabetics with CKD of the same stages. |

| | | |
|---|---|---|
| | News Title | **당뇨 앓는 신부전증 환자, 고혈압에 취약** |
| | News Article | 당뇨병을 앓고 있는 신부전증 환자가 고혈압에 더 취약한 것으로 나타나 주의가 필요하다. 인제대학교 일산백병원 신장내과 오세원 교수팀이 21개 병원에서 만성 신부전증을 앓고 있는 고혈압 환자 1천312명(당뇨 환자 439명 포함)을 대상으로 24시간 혈압을 측정해 분석했다. 그 결과 당뇨병을 동반한 환자가 아침과 야간에 혈압이 더 많이 상승하는 것으로 나타났다. ....중략.... 염분 섭취를 많이 할 경우 고혈압, 비만, 대사 증후군을 유발할 수 있어 특히 만성 신부전 환자에서 저염분 식이가 강조된다"며 "저염식이를 할 수 있는 방법은 국물이 있는 음식(국, 찌개, 면류)에서 국물은 먹지 않고 건더기만 건져 먹는 방법, 라면 등의 인스턴트 음식을 먹지 않는 방법, 외식을 줄이기, 김치 등을 싱겁게 담그는 방법, 간을 할 때 소금 대신 간장이나 고추장 혹은 허브를 이용하는 방법 등이 있다"고 조언했다. 이번 논문은 국제학술지 고혈압 연구(Hypertension Research) 최근호에 게재됐다. |