# 효과적인 한국어 교차언어 전송을 위한 특성 연구

윤태준°, 김태욱†
한양대학교 컴퓨터소프트웨어학부
tj1616@hanyang.ac.kr, kimtaeuk@hanyang.ac.kr

# Research on Features for Effective Cross-Lingual Transfer in Korean

Taejun Yun°, Taeuk Kim†
Department of Computer Science, Hanyang University

**요 약**

자원이 풍부한 언어를 사용하여 훈련된 모델을 만들고 해당 모델을 사용해 자원이 부족한 언어에 대해 전이 학습하는 방법인 교차언어 전송(Cross-Lingual Transfer)은 다국어 모델을 사용하여 특정한 언어에 맞는 모델을 만들 때 사용되는 일반적이고 효율적인 방법이다. 교차언어 전송의 성능은 서비스하는 언어와 전송 모델을 만들기 위한 훈련 데이터 언어에 따라 성능이 매우 다르므로 어떤 언어를 사용하여 학습할지 결정하는 단계는 효율적인 언어 서비스를 위해 매우 중요하다. 본 연구에서는 교차언어 전송을 위한 원천언어를 찾을 수 있는 특성이 무엇인지 회귀분석을 통해 탐구한다. 또한 교차언어전송에 용이한 원천 학습 언어를 찾는 기존의 방법론들 간의 비교를 통해 더 나은 방법을 도출해내고 한국어의 경우에 일반적으로 더 나은 원천 학습 언어를 찾을 수 있는 방법론을 도출한다.

주제어: 교차언어 전송, 다국어 언어모델

## 1. 서론

언어모델은 사전학습 언어모델의 등장[1] 이후로 많은 분야에서 널리 응용되고 있고 그 수요는 날이 갈수록 증가하고 있다. 다만 현재 자연어 처리분야의 대부분의 연구는 영어에 치중되어 있는데, 이는 영어 이외의 언어에 경우에 있어 자연어 처리 서비스를 제공하는데 큰 장애물이 되고 있다. 다국어를 처리 할 수 있는 모델인 다국어 사전학습 모델이 출현하였지만[2] 기존의 사전학습 언어모델을 원하는 작업에 맞게 미세조정(Fine-Tuning) 하는 과정이 필요한 현재의 패러다임에서는 다국어 모델을 사용하더라도 작업별 미세조정은 필수적이며, 미세조정을 위한 데이터가 부족한 언어들의 경우 여전히 서비스에 어려움이 있는것이 사실이다.

기존의 많은 연구들은 미세조정을 위한 데이터가 부족한 언어의 경우, 영어와 같은 고자원 언어(High-Resource Language)의 힘을 빌려 표적 작업(Target Task)에 대한 일반적인 지식을 학습하고 해당 지식을 실제 서비스하려고 하는 언어에 대해 전송하는 교차언어 전송(Cross-Lingual Transfer)을 통해 해결하곤 하였다. 교차언어 전송은 사전학습 언어모델의 출현 이전부터 일반적으로 자원이 부족한 언어에 대한 서비스를 위해 많이 사용됐던 방법이다. 하지만 교차언어 전송은 사전학습 언어모델의 출현 이후 더욱 주목받고 있는데, 그 이유는 사전학습 언어모델을 사용한 교차언어 전송의 성능이 강력하기 때문이다. 기존 연구[3]에 따르면 사전학습 언어모델을 활용한 교차언어 전송의 경우 영어로 표적 작업에 대해 학습한 모델을 사용한다. 이를 사용하여 서비스 하고자 하는 언어에 대해 적은 데이터만으로 추가 학습하는 퓨샷 교차언어 전송(Few-Shot Cross-Lingual Transfer)의 경우 100개의 추가 데이터만으로도 원천 학습 점수 (영어 점수) 대비 88% 정도의 성능을 기록하고 있으며, 심지어는 영어로 학습한 지식을 별다른 과정 없이 바로 전송하는 제로샷 교차언어 전송(Zero-Shot Cross-Lingual Transfer)의 경우에도 합리적인 성능을 보고하고 있다. 제로샷 교차언어전송은 별다른 훈련 과정 없이 영어와 같이 일반적으로 학습된 모델을 사용하여 작업에 대한 지식을 전송하기 때문에 컴퓨팅 자원적인 측면과 데이터 자원적인 측면에서 우수하며 실제로 사용 가능한 성능을 보여주고 있기 때문에 매우 유망한 방법으로 떠오르고 있다.

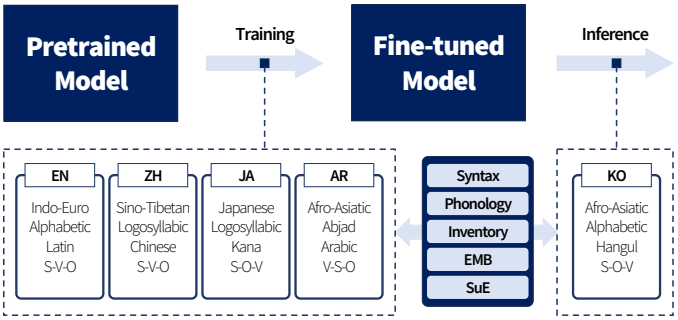


그림 1. 언어별 특징을 고려한 효율적인 교차언어 전송과정

한편, 교차언어 전송은 효율적이고 강력하지만 해당 방법을 기존 활용 방식 그대로 한국어와 같은 언어에 적용하기에는 여

† : 교신저자(Corresponding Author)

러 문제점이 있다. 다시 말해, 일반적으로 교차언어 전송에서는 영어를 원천언어로 활용하는 것이 보편적이지만, [4]에 따르면 영어는 유럽권 내의 언어에 대해서만 효율적인 교차언어 전송이 가능한 한계가 있는 것으로 보고되었다. 또한 교차언어 전송에 적합한 원천언어는 적용하고자 하는 작업에 따라 미세하게 다르다는 주장도 존재한다[3]. 위와 같은 이유로 여러 연구에서는 상대적으로 데이터 자원이 부족한 언어를 대상으로 한 교차언어 전송의 효율적인 원천언어를 찾기 위한 여러 가지 방법론들을 제시하고 있다.

본 연구에서는 기존 연구들의 교차언어 전송의 성능을 예측하고 원천언어를 효율적으로 찾는 방법론들을 통합하여 여러 가지 과제에 대해 확장하여 분석한다. 또한 기존의 교차언어 전송을 위한 효율적인 원천언어 선택 방법을 한국어에 적용하며 한국어에 맞는 원천언어 탐색 방법을 제시한다. 그림1은 문법, 음운, 임베딩 등 다양한 특징을 활용하여 표적언어에 맞는 원천언어를 선택하는 본 연구의 방법론을 간략하게 표현한다.

## 2. 관련 연구

자연어 처리 분야에서 사용되는 교차언어 전송은 일반적으로 자원이 상대적으로 풍부한 하나의 언어로 하나의 과제에 대하여 모델을 훈련한 후 해당 모델을 사용하여 다른 언어의 동일한 과제를 해결하는 경우를 통칭한다. 다국어 언어모델의 출현[2] 이후로 교차언어 전송의 강력한 성능이 보고되었고 많은 연구들이 이를 분석하며 다국어 언어모델의 어떤 점이 교차언어 전송을 강하게 만들었는지[5], 어떻게 교차언어 전송능력을 더욱 강하게 만들 수 있는지[6], 어떤 조건에서 교차언어전송이 잘 수행되는지[3, 4, 7]을 분석하였다. 본 연구에서는 어떤 조건에서 교차언어전송이 잘 이루어지는지 분석하며 이에 따라 본 절에서는 해당내용과 관련된 연구를 서술한다.

[3]의 연구에서는 영어를 원천언어로 사용한 교차언어 전송의 강력한 성능을 보고함과 동시에 표적언어(Target Language; 지식을 전송받는언어)가 영어와 언어적인 특성이 얼마나 비슷한지 또 해당 언어가 얼마나 다국어 언어모델의 사전학습에 사용되었는지를 교차언어 전송의 성공 여부를 판가름할 수 있는 중요한 요소로 주장하였다. 또한 [7] 연구에서는 품사 태깅(Part-Of-Speech Tagging) 작업에 대해 65개의 원천언어와 105개의 표적언어를 대상으로 WALS 데이터베이스를 사용하여 회귀분석을 실시하였고, 그 결과 원천언어와 표적언어의 사전학습 여부가 교차 학습의 성공 여부에 매우 중요하다고 밝혔다.

위와 같이 성공적인 교차 학습을 위한 특징을 연구하는 연구 이외에도 실제로 서비스하려고 하는 언어에 맞는 원천언어를 찾는 방법론들에 관한 연구도 여럿 진행되었다. 먼저 [8]은 Lang2Vec[9]를 사용하여 언어 간 전송 결과를 예측하는데 유리한 특성을 분석하였다. 또한 SuE[4]는 교차언어 전송의 성공 여부와 각 언어의 토큰화 과정을 연관 지어 토큰화가 덜 균일하게 일어나는 언어가 원천언어로서 적합한 언어라고 주장하였다. 이후 [10]의 연구에서는 WALS 데이터베이스를 포함한 데이터와 언어 임베딩, 각 언어별 Perplexity 등을 활용하여 회귀 분석을 통해 작업별로 언어에 따른 교차언어 전송 결과를 예측할 수 있는 방정식을 도출하였다. 마지막으로 [11] 연구에서는 언어별 임베딩을 분석하여 언어적인 특성과 연관 지었고 언어별 임베딩이 교차언어 전송 결과와 상관관계가 높다는 것을 증명하였다.

표 1. 회귀분석 결과

| Task | NER | | POS | | PI | |
|---|---|---|---|---|---|---|
| Feature | coef | P-val | coef | P-val | coef | P-val |
| Family | 3.0017 | 0.014 | 3.8778 | 0.000 | **1.0365** | 0.101 |
| Script Type | **7.4964** | 0.000 | **5.7542** | 0.000 | 0.2623 | 0.812 |
| Script | 5.7706 | 0.000 | 1.0321 | 0.000 | **1.0365** | 0.101 |
| SOV order | -1.4681 | 0.090 | 1.7632 | 0.000 | 0.9376 | 0.330 |

본 연구에서는 위의 소개된 방법들을 통합하여 분석하고 한국어에 적합한 원천언어를 선택하는 데 있어 최적인 방법을 탐색한다.

## 3. 실험 환경

### 3.1 실험 모델

논문의 분석에 사용된 모델은 인코더 기반의 언어모델인 XLM-Roberta[2] 이다. 해당 모델은 RoBERTa[12]의 다국어 버전으로 Comon Crawl 말뭉치를 활용하여 100개의 언어에 대해 사전학습 시킨 다국어 언어 모델이다. 본 연구에서는 XLM-Roberta의 Base 모델을 사용하였으며, 해당 모델은 다수의 작업에서 우수한 교차언어 전송 점수를 기록하고 있다.

### 3.2 실험 데이터셋

본 논문에서는 3가지 작업(NER,POS,PI)에 대한 교차언어 전송 점수를 측정하고 분석한다. 첫 번째로 NER(Named Entity Recognition)의 경우 다국어 데이터셋인 WikiANN 데이터셋을 사용하며 해당 데이터셋에서 한국어를 포함한 17개의 언어를 각각 원천언어와 표적언어로 하여 분석을 진행한다. POS(Part-Of-Speech Tagging) 데이터셋의 경우 Universal Dependencies 2.8[13]을 사용하며 해당 데이터셋에서 한국어를 포함하여 20개의 언어에 대한 분석을 진행한다. 마지막으로 PI(Paraphrase Identification)의 경우 마찬가지로 다국어 데이터셋인 PAWS-X[14] 데이터셋을 전부 활용하며 한국어를 포함한 총 7개의 언어에 대한 분석을 진행한다. 교차언어 전송

표 2. 효율적 원천언어 선택을 위한 특성탐색(전체)

| Task | NER | | POS | | PI | | AVG | |
|---|---|---|---|---|---|---|---|---|
| Feature / Metric | NDCG@3 | ADV | NDCG@3 | ADV | NDCG@3 | ADV | NDCG@3 | ADV |
| Syntax | 0.6545 | 2.48 | **0.7878** | 1.60 | **0.9379** | **2.01** | **0.7934** | 2.03 |
| Phonology | 0.6107 | -1.04 | 0.7048 | -0.36 | 0.6329 | 0.51 | 0.6495 | -0.30 |
| Inventory | 0.6588 | 3.17 | 0.6694 | -3.45 | 0.7504 | 1.13 | 0.6929 | 0.28 |
| S+P+I | 0.6316 | -1.22 | 0.7802 | 1.45 | 0.8675 | 1.35 | 0.7598 | 0.53 |
| EMB | **0.7580** | **4.68** | 0.7504 | **2.37** | 0.8655 | 1.24 | 0.7913 | **2.76** |
| SuE | 0.5108 | -0.73 | 0.3818 | -8.00 | 0.6914 | 0.44 | 0.5280 | -2.76 |

점수의 측정 시 NER과 POS의 경우 F1-Score를 사용하고 PI의 경우에는 정확도(Accuracy)를 사용한다.

## 4. 교차언어 전송 회귀분석

본 절에서는 언어 간 교차전송에 영향을 미치는 요인을 탐색하기 위해 회귀분석을 실시한다. 회귀분석은 여러 개의 독립변수에 대해 하나의 종속변수를 예측할 수 있는 식을 도출해내는 방법으로, 이 과정에서 각 독립변수가 종속변수에 얼마나 영향을 끼치는지를 의미하는 회귀 계수를 도출해 낼 수 있다.

기존 연구들[7, 10]은 각 언어의 고유한 특성을 독립변수로, 두 언어 사이의 교차언어 전송 점수를 종속변수로 하여 회귀분석을 통해 교차언어 전송점수를 예측하고 어떤 특성이 교차언어 전송에 영향을 많이 끼치는지 분석하였다. 논문의 회귀 분석에서 사용하는 독립변수는 [7]의 논문에서 사용하는 WALS 데이터베이스를 기반으로 작성된 언어별 특성을 사용하며, 회귀분석 방식은 OLS(Ordinary Least Square)를 사용한다.

사용하는 특성에 대해 좀 더 상세하게 설명하자면 1. 원천언어와 표적언어가 계통학적으로 동일한 어족에 속해있는지(Family Same), 2. 원천언어와 표적언어가 동일한 문자 체계(음소문자,표어 문자,아브자드 등)에 속하는지(Script Type Same), 3. 동일한 문자(라틴문자, 한자, 한글 등)를 사용하는지, 4. 문장 내 주어-동사-목적어의 순서가 동일한지(SOV Order Same)로 구성되어 있다. 한국어 특성 조사 결과의 경우 WALS 데이터베이스의 분석 결과에 국립국어원의 정보를 추가로 활용하여 일부 수정하였다.

표1의 결과는 모든 언어 쌍을 사용하여 도출 해낸 교차언어 전송점수를 종속변수로 하여 진행한 회귀분석의 결과를 나타낸다. 표 1의 실험 결과는 동일한 문자 체계(Script Type Same)가 교차언어 전송에 가장 중요한 것으로 보고하고 있다. 다만 PI 작업에서는 예외적으로 동일한 문자 체계에 대한 회귀계수가 가장 작게 기록되었는데, 이는 실험 간에 PI에 대한 회귀분석이 모든 변수에 대한 P-Value가 0.05 이상으로 통계적 검증과정을 통과하지 못하였기 때문에 해당 경우에 관해서는 추가적인 조사가 필요할 것으로 사료된다. 이에 대한 우리의 분석은 PI의 분석에 사용된 데이터셋이 포함하는 언어가 작아 교차언어 전송 사례가 적었고 그로 인해 회귀분석에 사용할 수 있는 데이터 수가 적어 정확한 회귀분석이 어려웠다고 판단하고 있다.

표 3. 한국어 작업 기준 최고의 원천언어

| Task | NER | | POS | | PI | |
|---|---|---|---|---|---|---|
| Rank | Lang. | F1 | Lang. | F1 | Lang. | Acc. |
| 1 | FI | 82.86 | EN | 66.49 | JA | 80.65 |
| 2 | EL | 82.66 | SV | 64.99 | FR | 79.7 |
| 3 | HE | 82.36 | JA | 64.93 | ZH | 79.25 |
| 4 | AR | 81.51 | TR | 64.39 | DE | 79.1 |
| 5 | TR | 81.08 | FI | 64.38 | ES | 78.4 |

## 5. 적절한 원천언어 선택을 위한 특성 탐색

이번 절에서는 우수한 원천언어를 찾기 위한 기존의 연구를 통합하여 분석한다. 본 실험에서는 교차언어 전송을 위한 원천언어를 찾는 기존의 방법론들을 동일한 조건에서 비교한다. 또한 기존의 방법론들을 사용하여 원천언어로 적합한 언어의 순위를 매겨 해당 순위가 실제 순위와 얼마나 비슷한지 NDCG@3 메트릭을 사용하여 측정한다. 또한 본 실험에서는 각 표적언어 별로 기존의 방법론을 사용하여 찾은 최적의 원천언어가 일반적으로 사용되는 원천언어(영어) 대비 얼마나 교차언어 전송 점수에서 이점(ADV)을 가져올 수 있는지를 계산한 후 평균 내어 기록한다. 본 실험에서는 1. Lang2Vec[9]의 유형론(Typological)적인 특성을 사용한 유사도 측정, 2. 언어별 임베딩의 코사인 유사도 측정, 3. SuE[4]를 사용한 원천언어 예측, 세 가지 방법을 사용하여 각 표적언어 별로 원천언어를 탐색한다.

본 실험에서는 먼저 Lang2Vec의 Syntax, Phonology, Inventory 특성 벡터를 사용하여 원천언어를 탐색한다. Lang2Vec은 언어별 특징을 조사하여 벡터화한 라이브러리로 그중 Syn-

표 4. 효율적 원천언어 선택을 위한 특성탐색(한국어)

| Task | NER | | POS | | PI | | AVG | |
|---|---|---|---|---|---|---|---|---|
| Feature / Metric | NDCG@3 | ADV | NDCG@3 | ADV | NDCG@3 | ADV | NDCG@3 | ADV |
| Syntax | 0.3971 | -0.62 | 0.5086 | -1.88 | **0.9552** | **3.17** | 0.6203 | 0.22 |
| Phonology | 0.4697 | 0.00 | **0.8145** | **0.00** | 0.3821 | 0.00 | 0.5554 | 0 |
| Inventory | 0.6509 | **6.52** | 0.7353 | -1.77 | 0.6841 | 2.47 | 0.6901 | **2.41** |
| S+P+I | 0.4917 | -0.62 | 0.7359 | -1.88 | 0.8655 | **3.17** | 0.6977 | 0.22 |
| EMB | **0.9109** | **6.52** | 0.4056 | -3.17 | 0.8324 | 3.02 | **0.7163** | 2.12 |
| SuE | 0.6509 | 4.78 | 0.3039 | -6.51 | 0.8655 | **3.17** | 0.6068 | 0.48 |

tax, Phonology, Inventory 특성 벡터는 Typological 한 특성에 해당한다. 그 후 해당 벡터들을 사용하여 벡터 간의 코사인 유사도(Cosine-Similarity)를 측정하고 유사도가 높은 언어끼리 교차언어 전송이 잘 될 것으로 간주한다. Syntax는 문법을 의미하며 언어의 문법을 WALS 데이터베이스를 사용하여 벡터화한 결과이다. Phonologycal또한 유사한 방식으로 WALS 데이터베이스를 사용하여 언어의 음운론적인 정보를 벡터화한 결과이다. Inventory는 언어의 소리 적인 정보와 연관된 특성으로 PHOIBLE 데이터베이스를 활용하여 추출된 벡터이다. Syntax+Phonology+Inventory에 해당하는 S+P+I의 경우 각각의 특성벡터를 병합하여 사용한다.

언어별 임베딩에 해당하는 EMB의 경우에는 언어별로 원천언어 미세조정 모델을 만들기 위한 훈련 데이터 중(예: NER-Wikiann) 1024개의 데이터를 활용하여 언어별 1024개의 임베딩을 구한 후 평균 내어 사용한다. 그 후 언어별 유사도를 비교할 때는 생성된 벡터 간의 코사인 유사도를 계산하여 유사한 임베딩을 가진 언어끼리의 교차언어 전송이 더 유리할 것으로 간주한다.

SuE는 언어별로 최적의 원천언어를 추천하지 않는다. 해당 방식에서는 일반적으로 토큰화가 불균일하게 일어나는 언어가 교차언어 전송에 유리하다고 주장한다. 따라서 해당 방식의 성능측정은 모든 언어에 동일한 순위로 원천언어가 선정된다 토큰화의 불균형은 각각 언어 말뭉치를 활용하여 토큰화 한 후 얼마나 다양한 길이의 토큰이 생성되는지를 통해 측정된다. 본 실험의 측정방식에서는 토큰의 불균형을 측정하기 위한 언어 말뭉치로 원천언어 미세조정 모델을 만드는 데 사용한 훈련 데이터를 사용한다. 그 후 언급된 방법들이 표적언어에 맞는 원천언어를 잘 찾을수 있는지 NDCG@3와 ADV를 활용하여 평가하였고 그 결과를 표 2에 기록하였다.

표2의 결과는 표적언어에 대한 원천언어를 찾는 방법으로 EMB와 Syntax 두 가지 방법이 NDCG@3와 ADV 두가지 측정에서 모두 우수한 성능을 기록하였다는 것을 나타낸다. Syntax와 EMB는 전반적으로 모든 작업에서 우수한 NDCG@3 값이 관측되었고 모든 작업에서 영어 대비 이점을 얻을 수 있는 언어를 추천하였다. 특히 임베딩 벡터는 언어별 1024개의 데이터만 있다면 구할 수 있으며 별다른 외부 지식(문법, 음운 등) 없이도 구할 수 있기 때문에 언어별 데이터를 구하기 힘든 언어나 특징을 조사하기 힘든 언어에도 적용할 수 있다는 장점이 있어 전반적으로 성능 면이나 효율성 측면에서 가장 우수한 방법으로 나타났다.

## 6. 한국어 교차언어 전송 분석

본 절에서 우리는 앞선 교차언어 전송에 대한 분석을 한국어에 적용함으로써 한국어에 맞는 원천언어를 효율적으로 탐색하는 방법에 대해 탐구한다. 각 분석에 앞서 우리는 본 절의 분석의 이해를 위해 한국어에 대한 실제 최적의 원천언어들을 제공한다. 표3의 결과는 각 작업별 한국어에 대한 최적의 원천언어 5개와 해당언어를 사용한 교차언어 전송점수를 나타낸다.[1] 기록된 숫자는 NER과 POS작업의 경우 해당 원천언어를 사용한 모델의 한국어 F1-Score, PI의 경우에는 한국어 정확도(Accuracy)를 의미 한다.

### 6.1 회귀분석 결과의 한국어 적용

앞선 4절의 결과를 토대로 본 절에서는 한국어 교차언어 전송에 영향을 끼치는 요인을 분석한다. 한국어가 표적언어인 경우 어떤 언어적인 특성이 영향을 끼치는지 분석하기 위해서는 한국어가 표적언어인 교차언어 전송점수 만을 활용한 회귀분석을 실행하고 어느 특성의 회귀계수가 큰지 분석하는 것이 직관적이다. 하지만 해당 실험 조건에서는 데이터의 수가 너무 적어 통계적으로 유의미한 회귀분석이 이루어 질 수 없다고 판단하였다. 따라서 본 연구에서는 한국어에 특화된 회귀분석은 별도로 진행하지 않고 대신 표1 의 결과가 한국어에 대해서 유효한지 분석하였다.

먼저 한국어의 특성을 나열하면 1. 한국어는 다른 언어와는

[1]표3의 약자 정보는 다음과 같다. AR: 아랍어, DE: 독일어, EL: 그리스어, ES: 스페인어, EN: 영어, FI: 핀란드어, FR: 프랑스어, HE: 히브리어, JA: 일본어, SV: 스웨덴어, TR: 튀르크어, ZH: 중국어.

표 5. 원천언어로서의 한국어 분석

| Task | NER | | POS | | PI | | AVG | |
|---|---|---|---|---|---|---|---|---|
| Feature / Metric | Pearson | Spearman | Pearson | Spearman | Pearson | Spearman | Pearson | Spearman |
| Syntax | 0.2378 | 0.2853 | -0.0696 | 0.1070 | -0.8182 | -0.8286 | -0.2167 | -0.1454 |
| Phonology | -0.0773 | -0.1817 | 0.4630 | 0.4714 | 0.7676 | 0.6377 | 0.3844 | 0.3092 |
| Inventory | -0.4666 | -0.5088 | -0.2094 | -0.2457 | 0.1819 | 0.4286 | -0.1647 | -0.1087 |
| EMB | 0.8775 | 0.8971 | 0.2239 | 0.3035 | -0.4300 | -0.6000 | 0.2238 | 0.2002 |

독립된 한국어족으로 분류되어 있고, 2. 문자 체계의 경우 영어와 동일한 음소문자(Alphabetic), 3. 독립적인 문자인 한글을 사용하며, 4. 주어-목적어-동사의 순서대로 어순이 구성되어 있다. 앞선 회귀분석의 결과는 동일한 문자 체계가 교차언어 전송에 긍정적인 영향을 끼친다는 것을 나타내고 있다. 실제로 표3의 결과에 따르면 NER에서 한국어에 대해 최적의 원천언어였던 그리스어(EL)와 핀란드어(FI)는 한국어와 동일한 음소문자에 해당하는 언어이고 POS에서 최적의 원천언어였던 영어(EN), 터키어(TR)모두 음소문자에 해당하는 언어로 밝혀졌다. 이를 통해 본 논문에서는 한국어가 표적언어인 경우에도 동일한 문자 체계가 교차언어 전송에 긍정적인 영향을 끼친다고 잠정적인 결론을 낼 수 있다.

### 6.2 한국어 원천언어 탐색방법 분석

5절은 적절한 원천언어를 찾기 위해 어떤 방법을 사용해야 할지 분석하고 있다. 그렇다면 한국어에 맞는 원천언어는 어떤 방식으로 찾는 것이 적합할까? 표4의 결과는 표적언어를 한국어로 제한한 실험의 결과 즉 한국어에 적합한 원천언어를 예측하는 방법에 대한 실험 결과이다. 전체적으로 NDCG의 관점에서는 임베딩이, ADV의 관점에서는 Inventory가 원천언어를 찾는 데 있어 우수하다고 밝혀졌다. 실제로 한국어와 문법(Syntax)이 가장 유사한 언어는 일본어(JA), 힌디어(HI)로 측정되었고, 음운론적으로(Phonologycal) 유사한 언어는 영어(EN)와 인도네시아어(ID)로 측정되었다. 소리적으로 유사한 언어(Inventory)는 핀란드어(FI)와 타이(TH)언어, 프랑스어(FR)로 측정되었고 세 가지 특징을 모두 사용하여 측정한 유사도는(S+P+I) 일본어가 가장 높게 측정되었다. 임베딩(EMB)의 경우 작업별로 다르게 측정되었는데 NER의 경우 핀란드어가, POS의 경우 인도네시아어가, PI의 경우 중국어(ZH)가 한국어와 가장 유사한 특징을 가지고 있다고 측정되었다. 임베딩은 실제로 NER에서 가장 우수한 원천언어인 핀란드어와 PI에서 3번째로 우수한 원천언어였던 중국어를 잘 예측하였다. Inventory 특성또한 NER에서 가장 우수한 원천언어인 핀란드어를 잘 식별하였으며 PI에서 우수한 원천언어였던 프랑스어 또한 한국어와 Inventory 벡터가 유사한 언어임으로 드러났다. 실험 결과를 바탕으로 본 논문은 한국어의 경우에도 언어별 임베딩 벡터를 사용하여 코사인 유사도를 계산하는 방식이 유효한 방법이며 Inventory 특성을 추가로 사용할 수 있음을 입증하였다.

### 6.3 원천언어로서의 한국어 분석

앞선 [3]의 연구에서는 가장 자원이 풍부한 언어인 영어를 원천언어로 하여 얼마나 영어로부터 잘 학습할 수 있는지와 연관된 특성을 분석하였다. 한국어는 영어와 중국어와 같은 언어에 비해서는 자원이 부족한 언어이지만 스와힐라이어(SW)와 우두르(UR)어 같은 언어에 비해서는 상대적으로 더 자원이 풍부한 언어이다. 본 절에서는 사용하는 원천언어를 한국어로 치환하여 교차언어 전송을 평가하여 한국어를 원천언어로 훈련된 모델이 어느 언어에 전송이 용이한지 또 그러한 전송 결과는 언어의 어떤 특성과 상관관계가 있는지 분석한다. 해당 실험에 사용된 모델은 XLM-RoBERTa-Base이다.

표 5의 결과는 [3]의 분석의 원천언어를 한국어로 치환한 실험의 결과이다. 실험은 한국어를 원천언어로 하여 각 언어로 교차언어 전송된 결과와 한국어와 해당 언어와의 언어 특성 차이 간의 상관계수를 측정한다. 비록 NER 작업에서 EMB와의 높은 양의 상관관계, PI 작업에서 문법적 특성에서 음의 상관관계를 찾을 수 있었지만 모든 작업에서 일반적으로 한국어를 원천언어로 사용한 교차전송 정도와 상관있는 지표를 찾지는 못하였다. 이와 같은 결과에 대하여서는, 한국어 데이터가 더 풍부해지 보다 많은 표적언어 및 표적작업의 조합을 고려할 수 있는 환경이 조성된다면 보다 나은 일반화가 가능하리라 기대하며, 관련 내용을 후속 연구로 수행하고자 한다.

## 7. 결론

본 연구에서는 효율적인 교차언어 전송을 위한 원천언어 선택에 대한 기존 방법론들을 비교한다. 해당 분석의 결과로 기존의 존재하는 원천언어를 선택하는 방법론들 중 일반적으로 사용할 수 있는 방법을 탐색하며, 이를 한국어에 최적화시켜 한국어에 해당하는 최적의 원천언어를 추천할 수 있는 방법을 추천하였다.

## 감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1F1A1074674).

## 참고문헌

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019. [Online]. Available: https://aclanthology.org/N19-1423

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. [Online]. Available: https://aclanthology.org/2020.acl-main.747

[3] A. Lauscher, V. Ravishankar, I. Vulić, and G. Glavaš, "From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020. [Online]. Available: https://aclanthology.org/2020.emnlp-main.363

[4] O. Pelloni, A. Shaitarova, and T. Samardzic, "Subword evenness (SuE) as a predictor of cross-lingual transfer to low-resource languages," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022. [Online]. Available: https://aclanthology.org/2022.emnlp-main.503

[5] K. Karthikeyan, Z. Wang, S. Mayhew, and D. Roth, "Cross-lingual ability of multilingual bert: An empirical study," *International Conference on Learning Representations*, 2020.

[6] H. Yang, H. Chen, H. Zhou, and L. Li, "Enhancing cross-lingual transfer by manifold mixup," *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=OjPmfr9GkVv

[7] W. de Vries, M. Wieling, and M. Nissim, "Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.529

[8] Y.-H. Lin, C.-Y. Chen, J. Lee, Z. Li, Y. Zhang, M. Xia, S. Rijhwani, J. He, Z. Zhang, X. Ma, A. Anastasopoulos, P. Littell, and G. Neubig, "Choosing transfer languages for cross-lingual learning," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. [Online]. Available: https://aclanthology.org/P19-1301

[9] P. Littell, D. R. Mortensen, K. Lin, K. Kairis, C. Turner, and L. Levin, "URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2017. [Online]. Available: https://aclanthology.org/E17-2002

[10] B. Muller, D. Gupta, J.-P. Fauconnier, S. Patwardhan, D. Vandyke, and S. Agarwal, "Languages you know influence those you learn: Impact of language characteristics on multi-lingual text-to-text transfer," *Proceedings of The 1st Transfer Learning for Natural Language Processing Workshop*, 2023. [Online]. Available: https://proceedings.mlr.press/v203/muller23a.html

[11] P. Lin, C. Hu, Z. Zhang, A. F. T. Martins, and H. Schütze, "mplm-sim: Unveiling better cross-lingual similarity and transfer in multilingual pretrained language models," 2023.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[13] D. Zeman, J. Nivre *et al.*, "Universal dependencies 2.8.1," 2021. [Online]. Available: http://hdl.handle.net/11234/1-3687

[14] Y. Yang, Y. Zhang, C. Tar, and J. Baldridge, "PAWS-X: A cross-lingual adversarial dataset for paraphrase identification," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019. [Online]. Available: https://aclanthology.org/D19-1382