# Continual learning을 이용한 한국어 상호참조해결의 도메인 적응

최요한*O 조경빈* 이창기* 류지희** 임준호**
*강원대학교 빅데이터메디컬융합학과, **한국전자통신연구원
{choiyohan, jkb5570, leeck}@kangwon.ac.kr, {chrisjihee, joonho.lim}@etri.re.kr

# Domain adaptation of Korean coreference resolution using continual learning

Yohan Choi*O Kyengbin Jo* Changki Lee* Jihee Ryu** Joonho Lim**
*Department of Big Data Medical Convergence, Kangwon National University
**Electronics and Telecommunications Research Institute

**요 약**

상호참조해결은 문서에서 명사, 대명사, 명사구 등의 멘션 후보를 식별하고 동일한 개체를 의미하는 멘션들을 찾아 그룹화하는 태스크이다. 딥러닝 기반의 한국어 상호참조해결 연구들에서는 BERT를 이용하여 단어의 문맥 표현을 얻은 후 멘션 탐지와 상호참조해결을 동시에 수행하는 End-to-End 모델이 주로 연구가 되었으며, 최근에는 스팬 표현을 사용하지 않고 시작과 끝 표현식을 통해 상호참조해결을 빠르게 수행하는 Start-to-End 방식의 한국어 상호참조해결 모델이 연구되었다. 최근에 한국어 상호참조해결을 위해 구축된 ETRI 데이터셋은 WIKI, QA, CONVERSATION 등 다양한 도메인으로 이루어져 있으며, 신규 도메인의 데이터가 추가될 경우 신규 데이터가 추가된 전체 학습데이터로 모델을 다시 학습해야 하며, 이때 많은 시간이 걸리는 문제가 있다. 본 논문에서는 이러한 상호참조해결 모델의 도메인 적응에 Continual learning을 적용해 각기 다른 도메인의 데이터로 모델을 학습 시킬 때 이전에 학습했던 정보를 망각하는 Catastrophic forgetting 현상을 억제할 수 있음을 보인다. 또한, Continual learning의 성능 향상을 위해 2가지 Transfer Techniques을 함께 적용한 실험을 진행한다. 실험 결과, 본 논문에서 제안한 모델이 베이스라인 모델보다 개발 셋에서 3.6%p, 테스트 셋에서 2.1%p의 성능 향상을 보였다.

주제어: Continual learning, Catastrophic forgetting, 상호참조해결

## 1. 서론

상호참조해결(Coreference resolution)은 명사, 대명사, 명사구 등의 멘션(mention) 후보를 식별하고, 동일한 개체(entity)를 의미하는 멘션들을 찾아 그룹화(clustering)하는 자연어처리 태스크이다. 딥러닝 기반의 한국어 상호참조해결 연구들에서는 BERT를 이용하여 단어의 문맥 표현을 얻은 후 멘션 탐지와 상호참조해결을 동시에 수행하는 End-to-End 모델[1]이 주로 연구가 되었으며, 최근에는 모델의 계산 복잡도를 줄이기 위해 스팬 표현을 사용하지 않고 멘션의 시작과 끝 표현을 사용하는 Start-to-End 한국어 상호참조해결 모델[2]이 연구되었다.

최근에 한국어 상호참조해결을 위해 구축된 ETRI 데이터셋은 WIKI, QA, CONVERSATION 등 다양한 도메인으로 이루어져 있으며, 신규 도메인의 데이터가 추가될 경우 신규 데이터가 추가된 전체 학습데이터로 모델을 다시 학습해야 하며, 이때 많은 시간이 걸리는 문제가 있다. 이러한 문제를 해결하기 위해서 본 논문에서는 한국어 상호참조해결의 도메인 적응에 Continual learning을 적용한다.

Continual learning은 Transfer learning과 유사하게 기존의 학습 내용을 유지하면서 새로운 정보만 추가로 학습하기 위해 개발된 알고리즘으로, Transfer learning에서 문제가 되는 Catastrophic forgetting(이전 태스크에 대한 정보의 망각) 문제를 방지하는 것에 중점을 두고 연구되고 있다.

본 논문에서는 Longformer[6]를 이용한 Start-to-End 한국어 상호참조해결 모델의 도메인 적응을 위해 Continual learning 기술 중에서 Regularization 방식의 Elastic Weight Consolidation[3]와 Incremental Moment Matching[4] 알고리즘을 각각 적용하고 이 둘을 같이 적용했을 때의 성능을 비교하고, 추가로 L2-Transfer와 Drop-Transfer 기술을 함께 적용한다. 상호참조해결 모델의 도메인 적응 실험은 3가지 다른 도메인의 데이터들을 WIKI, QA, CONVERSATION 도메인 순으로 모델을 학습시킨 후, 최종 모델에서 각 도메인의 성능을 측정하여 Catastrophic forgetting 현상을 어느 정도 억제하는지 비교한다. 실험 결과, 베이스라인(Transfer learning) 대비 개발 셋에서 3.6%p, 테스트 셋에서 2.1%p가량의 성능이 향상되었다.

## 2. 관련 연구

**Transfer learning** 이전 태스크로 학습된 모델을 새로운 태스크에 사용하는 방법이다. 이를 위해선 학습된 모델이 필요하며, 본 논문에서는 WIKI, QA, CONVERSATION 순의 3개 도메인을 통해 Transfer learning을 구현한다.

**Continual learning** 기존의 학습 내용을 유지하면서 새로운 정보를 추가로 학습하기 위한 것으로, 주로 Catastrophic forgetting이라 불리는 이전 태스크에 대한

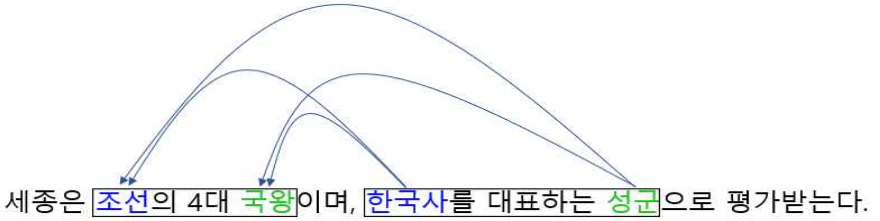


그림 1 Start-to-End 한국어 상호참조해결 모델의 수행 방식, '조선의 4대 국왕'과 '한국사를 대표하는 성군'의 시작-시작, 시작-끝, 끝-시작, 끝-끝 표현의 4가지 연산으로 상호참조해결을 수행한다.

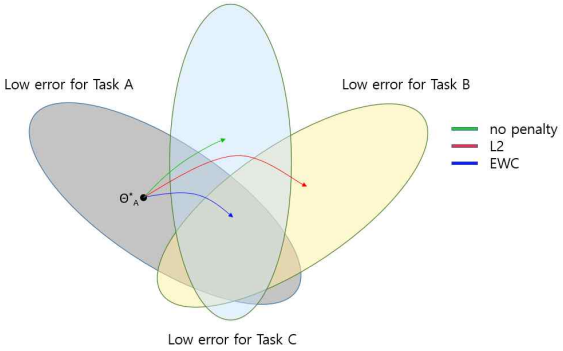


그림 2 EWC의 예시 그림, 태스크 A를 학습한 후 모델의 가중치가 $\theta^*_A$라면, EWC는 이후의 태스크 B, C를 학습하며 가 가중치가 태스크 A, B, C에 대해 적은 오차를 가지는 분포로 유도한다.

정보의 망각을 방지하기 위한 것에 중점을 두고 연구가 진행되고 있다. Continual learning은 크게 Regularization, Distillation, Structure의 3가지 방향으로 발전되었는데 본 논문에서는 Regularization 방식의 Elastic Weight Consolidation(EWC) 모델과 그의 변형인 Incremental Moment Matching(IMM) 모델을 사용한다.

## 3. Continual learning을 적용한 도메인 적응 기술

본 논문에서는 한국어 상호참조해결 모델의 도메인 적응 성능 향상을 위해 Continual learning을 적용한 모델을 제안한다. 이를 위해, 기존의 상호참조해결 모델[2]에 Fisher Information Matrix를 손실 함수의 정규화 항으로 추가하는 EWC와 weight를 병합하는 데 이용하는 IMM을 각각 적용하고 이 둘을 같이 적용하는 모델의 성능을 비교한다.

**Start-to-End Coreference Resolution** 기존 End-to-End 상호참조해결 모델에서 계산하는 스팬 표현을 적용하지 않아 학습 및 실행속도가 빠른 모델이며, 한국어 특성에 맞게 추가로 개체명 자질과 의존 구문분석 자질이 추가되었다. Start-to-End 상호참조해결 모델은 End-to-End 모델과 같이 멘션 스코어를 이용하여 상위 k개의 멘션 후보를 추출하고 스팬 표현을 사용하지 않아 계산량이 적기 때문에 추출한 k개의 모든 멘션 후보들에 대해 선행사 스코어를 계산한다. 선행사 스코어도 시작-시작, 시작-끝, 끝-시작, 끝-끝의 시작과 끝 표현식을 계산한다.

**Elastic Weight Consolidation(EWC)** Regularization 방식의 대표적인 Continual learning 모델로, 이전 태스크를 학습하는데 있어 중요한 weight와 그렇지 않은 weight를 계산해 이를 학습에 이용한다. Fisher Information Matrix를 이용해 weight간 중요도를 계산하며, 새로운 태스크를 학습할 때는 이전 태스크의 학습에서 중요한 weight의 변화를 최대한 억제하도록 유도한다.

$$L(\theta) = L_B(\theta) + \sum_i \frac{\lambda}{2} F_i(\theta_i - \theta^*_{A,i})^2 \quad (1)$$

식 (1)은 EWC의 주요 손실 함수이다. 기존 태스크와 새로운 태스크를 비교하고 이를 손실 함수의 정규화 항으로 추가한다. Fisher Information Matrix를 나타내는 F라는 함수의 $\theta_i$와 $\theta^*_{A,i}$로 이전 태스크를 학습할 때 중요한 weight를 계산하며, $\lambda$는 이전 태스크가 얼마나 중요한지를 조절하는 하이퍼 파라미터이다.

**Incremental Moment Matching(IMM)** EWC의 변형이며 여러 태스크로 모델을 각각 학습하고 이를 저장한 후 Mean-IMM과 Mode-IMM 2가지 방법으로 weight를 병합하는 모델이다. Mean-IMM은 단순히 각 태스크의 weight를 가중 평균하는 것이고, Mode-IMM은 Mean-IMM과 달리 공분산 행렬을 이용해 Gaussian Mixture를 수행한다.

$$p_{1:K} \equiv p(\theta \mid X_1, \cdots, X_K, y_1, \cdots, y_K) \approx q_{1:K} \equiv q(\theta \mid \mu_{1:K}, \Sigma_{1:K}) \quad (2)$$

$$p_k \equiv p(\theta \mid X_k, y_k) \approx q_k \equiv q(\theta \mid \mu_k, \Sigma_k) \quad (3)$$

IMM은 K개의 태스크에 대해 각 k번째 태스크의 가중치로 $q_{1:K}$의 최적 매개변수를 찾는 방식으로 수행된다.

식 (2)의 $q_{1:K}$는 전체 태스크에 대한 실제 가중치 분포인 $p_{1:K}$의 근사치이고, 식 (3)의 $q_k$는 k번째 태스크에 대한 훈련 데이터의 실제 가중치인 $p_k$의 근사치이다.

IMM은 여기에 성능 향상을 위해 L2-Transfer와 Drop-Transfer라는 두 가지 기술을 추가한다. L2-Transfer는 IMM의 손실 함수에 이전 태스크와 새로운 태스크 사이의 거리 정규화 항을 추가하며 Drop-Transfer는 드롭아웃의 변형으로 드롭아웃으로 인해 Off된 노드를 0 대신 이전 태스크의 weight로 대체하는 것이다[4].

## 4. 실험 및 결과

본 논문에서 사용한 데이터는 ETRI 상호참조해결 데이터 중의 WIKI, 질의응답, 대화문 도메인 데이터 셋이며 성능 측정을 위해 중심어 경계(head boundary)를 기준으로 muc, $b^3$, ceaf-e, CoNLL-F1[4]을 사용하였다. 상호참조해결의 도메인 적응 성능 비교를 위해 WIKI, QA, CONVERSATION 도메인 순서로 Continual learning을 적용한 뒤에, 3개 도메인의 성능을 측정하였다. 또한, Continual learning과의 비교 실험을 위해, 베이스라인 모델로는 기존 태스크를 학습시킨 모델을 가져와 새로운 태스크를 학습할 때 사용하는 Transfer learning을 사용하였다. 실험에 사용한 데이터 셋의 문서 수는 다음과 같다.

표 1 사용한 데이터 셋

| | Train | Dev | Test |
|---|---|---|---|
| WIKI | 1,296 | 50 | 50 |
| QA | 2,819 | 645 | 570 |
| CONVERSATION | 1,000 | 150 | 150 |

본 논문에서의 상호참조해결 모델 학습을 위한 하이퍼 파라미터는 다음과 같으며, 실험에 사용한 GPU는 RTX TITAN이다.

표 2 상호참조해결 하이퍼 파라미터

| | |
|---|---|
| 학습률 | 7e-06 |
| 드롭아웃 | 0.3 |
| Hidden layer 차원 수 | 3072 |
| 최대 스팬 길이 | 70 |
| k (최대 멘션 후보 수) | 전체 시퀀스 길이 40% |
| 옵티마이저 | AdamW |

표 3 Continual learning 하이퍼 파라미터

| | |
|---|---|
| Importance | 10000 |
| $\alpha$ | 0.6 |
| $\lambda$ | 1e-04 |

표 3은 Continual learning에 적용된 하이퍼 파라미터이다. Importance는 EWC의 Fisher Information Matrix에 쓰이고, $\alpha$는 Mean-IMM의 가중 평균에서의 가중치와 Mode-IMM의 Gaussian Mixture에서의 가중치 역할을 하는 하이퍼 파라미터이다. $\lambda$는 L2-Transfer Term을 조절하는 하이퍼 파라미터이다.

실험에 쓰인 한국어 상호참조해결 모델은 Longformer 기반의 Start-to-End 모델이며, [1]에서와 마찬가지로 후행 언어라는 한국어 특성에 맞게 Start-to-End 상호참조해결모델의 목적 함수를 재구성하였다.

표 4 전체 도메인 통합 데이터로 학습한 모델의 성능

| DEV | | | | |
|---|---|---|---|---|
| | muc | $b^3$ | ceaf-e | F1 |
| Upper bound | 73.1 | 68.3 | 68.9 | 70.1 |
| TEST | | | | |
| | muc | $b^3$ | ceaf-e | F1 |
| Upper bound | 72.9 | 68.0 | 67.9 | 69.6 |

표 4는 WIKI, QA, COVERSATION 도메인의 데이터를 통합하고 이를 동시에 모델에 학습시켰을 때의 성능으로 상호참조해결의 도메인 적응 실험에서의 성능의 최고 목표 수치(Upper bound)에 해당한다. 표의 muc, $b^3$, ceaf-e, CoNLL-F1 스코어는 3개의 도메인으로 각각 측정한 성능의 산술 평균이다.

표 5 EWC에 IMM을 적용한 모델 성능 비교

| DEV | | | | |
|---|---|---|---|---|
| | muc | $b^3$ | ceaf-e | F1 |
| Baseline | 64.9 | 59.9 | 59.6 | 61.4 |
| Mean-IMM | 66.4 | 62.0 | 62.3 | 63.5 |
| Mode-IMM | 66.8 | 62.3 | 62.4 | 63.8 |
| EWC | 66.3 | 61.4 | 61.1 | 62.9 |
| EWC+Mean-IMM | 68.0 | 62.2 | 62.4 | **64.2** |
| EWC+Mode-IMM | 66.8 | 61.9 | 62.0 | 63.5 |
| TEST | | | | |
| | muc | $b^3$ | ceaf-e | F1 |
| Baseline | 66.3 | 60.8 | 61.0 | 62.7 |
| Mean-IMM | 67.6 | 63.0 | 62.9 | 64.5 |
| Mode-IMM | 68.0 | 63.4 | 62.8 | 64.7 |
| EWC | 67.5 | 62.1 | 61.3 | 63.6 |
| EWC+Mean-IMM | 67.8 | 63.1 | 63.1 | **64.7** |
| EWC+Mode-IMM | 67.2 | 62.0 | 61.8 | 63.6 |

표 5는 베이스라인(Transfer learning), IMM, EWC, EWC에 IMM을 추가한 모델 간의 상호참조해결의 도메인 적응 성능을 비교한 것이다. 상호참조해결의 도메인 적응 성능 비교를 위해 WIKI, QA, CONVERSATION 도메인 순서로 Continual learning을 적용한 뒤에, 3개 도메인의 성능을 평균한 값이다. 개발 셋에서는 Mode-IMM과 EWC에 Mean-IMM을 적용했을 때 성능이 가장 좋으며, 테스트 셋에서는 EWC에 Mean-IMM을 적용했을 때 성능이 가장 좋음을 알 수 있다.

표 6 Transfer Techniques을 적용한 모델 성능 비교

| DEV | | | | |
|---|---|---|---|---|
| | muc | $b^3$ | ceaf-e | F1 |
| Baseline | 64.9 | 59.9 | 59.6 | 61.4 |
| Mean-IMM | 66.4 | 62.0 | 62.3 | 63.5 |
| Mode-IMM | 66.8 | 62.3 | 62.4 | 63.8 |
| Mean-IMM + L2-Transfer | 63.0 | 57.8 | 57.4 | 59.4 |
| Mode-IMM + L2-Transfer | 66.4 | 61.2 | 60.6 | 62.7 |
| Mean-IMM + Drop-Transfer | 66.5 | 61.9 | 62.3 | 63.5 |
| Mode-IMM + Drop-Transfer | 67.5 | 62.2 | 62.3 | **64.0** |
| TEST | | | | |
| | muc | $b^3$ | ceaf-e | F1 |
| Baseline | 66.3 | 60.8 | 61.0 | 62.7 |
| Mean-IMM | 67.6 | 63.0 | 62.9 | 64.5 |
| Mode-IMM | 68.0 | 63.4 | 62.8 | 64.7 |
| Mean-IMM + L2-Transfer | 65.7 | 60.1 | 59.4 | 61.7 |
| Mode-IMM + L2-Transfer | 67.0 | 62.0 | 61.8 | 63.6 |
| Mean-IMM + Drop-Transfer | 66.9 | 62.6 | 63.2 | 64.2 |
| Mode-IMM + Drop-Transfer | 68.4 | 62.9 | 63.1 | **64.8** |

표 6는 Mean-IMM과 Mode-IMM 모델에 성능 향상을 위해 L2-Transfer와 Drop-Transfer를 각각 적용한 모델 간의 성능 비교이다. 개발 셋과 테스트 셋 모두 Mode-IMM에 Drop-Transfer를 적용했을 때 가장 좋은 성능을 보임을 알 수 있으며, 베이스라인에 비해 약 2.1%p ~ 3.6%p의 성능 향상을 보였다.

## 5. 결론

본 논문에서는 Start-to-End 방식의 상호참조해결 모델에 도메인 적응 기술인 Continual learning 알고리즘 중에서 Regularization 방식에 해당하는 EWC와 IMM에 2가지 Transfer Techniques를 적용하였다. 실험 결과 Mode-IMM에 Drop-Transfer를 적용하였을 때 이전 태스크에 대한 정보를 망각하는 Catastrophic forgetting이 가장 잘 억제되는 것을 보였다. 구체적으로는 Mode-IMM에 Drop-Transfer를 더한 모델이 베이스라인 대비 개발 셋에서 3.6%p, 테스트 셋에서 2.1%p의 성능 향상을 보였다. 향후에는 본 본문에서 성능 향상이 저조한 Transfer Techniques를 수정하여 성능을 더욱 향상 시킬 예정이다.

## 감사의 글

이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.2013-2-00131, 휴먼 지식증강 서비스를 위한 지능 진화형 Wise QA 플랫폼 기술 개발).

## 참고문헌

[1] 김기훈, 박천음, 이창기, 김현기, "BERT기반 End-to-end 신경망을 이용한 한국어 상호참조해결", 정보과학회논문지, 제47권, 제10호, 2020.

[2] 조경빈, 정영준, 최요한, 이창기, 류지희, 임준호, “스팬 표현을 사용하지 않는 어절 단위의 한국어 End-to-End 상호참조해결”, 한국정보과학회 학술발표논문집, 2022, p.431-433.

[3] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, Raia Hadsell, “Overcoming catastrophic forgetting in neural networks”,Proceedings of the National Academy of Sciences, Volume 114, Issue 13, p.3521-3526.

[4] Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, Byoung-Tak Zhang, “Overcoming Catastrophic Forgetting by Incremental Moment Matching”, arXiv eprint arXiv:1703.08475, 2017.

[5] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, "Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules," Computational Linguistics, Volume 39, Issue 4, 2013.

[6] Iz Beltagy, Matthew E. Peters, Arman Cohan, “Longformer: The Long-Document Transformer”, arXiv eprint arXiv:2004.05150, 2020.