관계 추출 및 지식베이스 확장을 위한 반복 학습 시스템 설계

정용빈○, 남상하, 김지성, 이민호, 최기선
한국과학기술원, 시멘틱웹첨단연구센터
{kuonom, nam.sangha, jiseong, pathmaker, kschoi}@kaist.ac.kr

Iterative learning system design for relation extraction and knowledge base population

Yong-Bin Jeong○, Sang-Ha Nam, Ji-Seong Kim, Min-Ho Lee, Key-Sun Choi
KAIST, Semantic Web Research Center

요 약

관계추출기의 학습을 위해서는 많은 학습 데이터가 필요한데, 사람이 모으게 되면 많은 비용이 필요하여 원격 지도 학습을 이용한 데이터 수집이 많은 연구에서 사용되고 있다. 원격 지도 학습은 지식베이스를 기반으로 학습 데이터를 자동으로 만들어 내는 방식이기에 비용이 거의 들지 않지만, 지식베이스의 질과 양에 영향을 받는다. 본 연구는 원격 지도 학습을 기본으로 관계추출기의 성능을 향상 시키고, 지식베이스를 확장하는 방안으로 반복학습을 제안한다. 실험을 적은 비용으로 빠르게 진행하기 위해 반복학습을 자동화 하는 시스템을 설계하여 실험을 하였고, 이 시스템으로 관계추출기의 성능이 향상 될 수 있는 가능성을 보였으며, 반복학습을 통한 지식베이스의 확장 방안을 제시한다.

주제어: Relation Extraction, Distant Supervision, knowledge base population, Iterative Learning

1. 서론

지식베이스(knowledge base)는 일반 상식부터 전문 지식까지 사실 및 규칙 등이 저장되어있는 데이터베이스이다. 지식베이스는 질의응답 등 여러 분야에서 사용될 수 있기 때문에 DBpedia[1], Yago[2], Freebase[3] 와 같이 다양한 곳에서 구축되었고, 많은 연구가 이루어지고 있다. 관계추출은 자연언어처리에서 기본적인 과업 중 하나이고, 지식베이스를 구축하는데 필수적인 요소로서, 그림 1과 같이 일반적인 문장에서 두 개의 개체가 관계추출의 대상으로 지정되면, 문장에서의 두 개체간의 관계를 정의하는 과업이다. 본 연구에서는 한국어 지식베이스 중 하나인 Kbox[4]를 기반으로 하여 반복 학습(iterative learning)을 이용해 관계추출기의 성능을 향상시키고, 지식베이스를 확장하는 방안을 제시한다. 여기서 Kbox의 기본 지식단위는 트리플인데, <페이스북, leader, 마크 저커버그> 와 같이 두 개의 개체(페이스북, 마크 저커버그)와 그 관계(leader)를 나타내는 것이다.

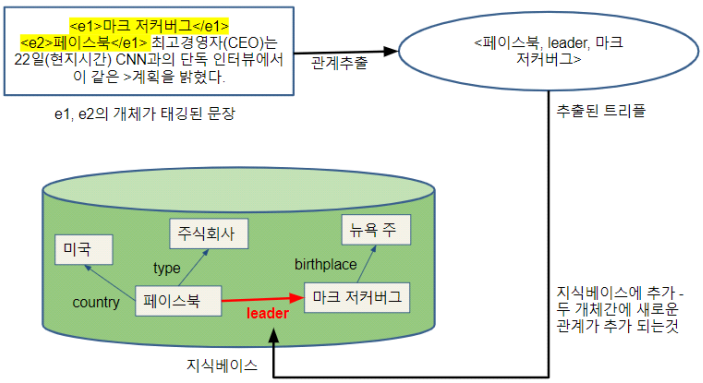

그림 1 관계추출 후 지식베이스에 지식이 추가되는 과정

본 연구의 반복학습에서 사용된 관계추출기를 학습시키는 데이터는 Kbox를 기반으로 한국어 위키피디아 데이터에 원격 지도 학습(distant supervision)[5]을 적용해 얻어진다. 그림 1과 같이 학습된 관계추출기를 이용해 문서들에서 관계를 추출하여 지식베이스의 단위 지식인 트리플(triple)을 생성해 지식베이스에 추가하여 지식베이스를 확장시킨다. 여기서, 확장된 지식베이스를 가지고 원격 지도 학습을 통해 다시 학습 데이터를 모으게 되면, 처음에 관계추출기를 학습했을 때 보다 더 많은 학습 데이터를 얻을 수 있게 되고, 따라서 관계추출기의 성능이 향상된다. 그 후에, 성능이 향상된 관계추출기를 이용해 관계를 추출하게 되면 지식베이스가 더 늘어날

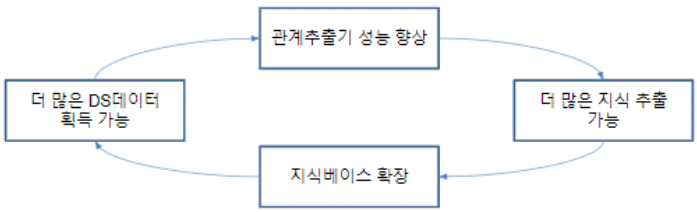

그림 2 반복학습 선순환의 기본 가정

것이다. 이를 반복하여 관계추출기의 성능을 향상시키고 지식베이스를 확장하는 것이 반복학습이다.

이와 같은 반복학습에서는 문제점이 될 수 있는 것이 몇 가지 존재하는데, 그 중 하나는 반복학습이 진행됨에 따라 원격 지도 학습으로 얻은 데이터의 질이 오히려 더 낮아질 수 있다는 점이다. 지식을 추출하는 관계추출기가 잘못된 지식을 지식베이스에 저장하게 되면 원격 지도 학습도 부정확한 학습데이터를 모을 가능성이 높아지고, 그로 인하여 오히려 성능이 나빠질 수도 있다. 그렇기 때문에, 지식 검증을 통해 올바른 지식만 골라서 지식베이스에 추가하는 방법을 제안하는 것과, 관계추출기가 얼마만큼의 오류까지 견디고 성능이 향상될 수 있는지를 확인 하는 것도 본 연구의 목표중 하나이다.

지식 검증이 반복학습에 어떤 영향을 주는지 확인하기 위해서는 다양한 지식 검증 알고리즘을 적용하여 많은 실험을 할 필요가 있다. 그래서 본 연구에서는 반복학습을 자동화 하는 시스템을 설계하여 실험 과정에서 사람의 노력이 적게 들도록 하였다.

2. 관련 연구

반복학습과 비슷하게 진행되는 연구로는 Never ending language learning(NELL)[6] 이 있다. NELL에서는 주어진 온톨로지를 확장하기 위해 웹에서 요소들을 추출하고 추출을 더 잘하기 위해 다시 학습하는 것을 반복한다. 그리고 반복적인 학습으로 관계추출의 성능을 향상시키는 연구로는 ARNOR[7]가 있다. ARNOR는 적더라도 신뢰있는 데이터에서 시작해서 반복적으로 원격 지도 학습에서 얻는 데이터를 패턴에 따라 검증하여 검증된 데이터만 학습에 사용한다. 반복마다 패턴이 늘어 검증된 데이터의 양을 증가시킴으로써 관계추출기의 학습에 필요한 데이터가 좀 더 질이 좋은 데이터로 구성되도록 하는 방법이다. 이 이외에도 원격지도학습의 노이즈를 해소하기 위한 강화학습[8]이나 GAN[9] 등 많은 연구가 관계추출기의 성능 향상을 위해 이루어지고 있다.

3. 반복 학습 System

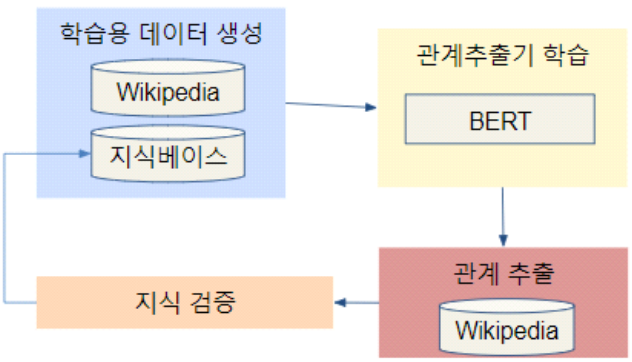

그림 3 반복학습

그림 3과 같은 반복학습을 위해서는 그림 4와 같이 개체 연결(entity linking)이 완료된 데이터, 초기 시드로 사용될 지식베이스, 관계추출기 모델이 필요하다. 그 뒤엔, 그림3 과 같이 반복학습을 진행하게 되는데, 첫 번째는 관계추출기를 학습시키기 위한 학습용 데이터 생성이다. 그 후에는 생성된 데이터로 관계추출기를 학습시키고, 학습된 관계추출기로 다시 문장에서 관계추출을 하여 트리플을 생성한다. 그 뒤, 생성된 트리플들을 검증하여 올바른 트리플이라 판단된 것을 지식베이스에 추가한다. 이 과정을 반복하여 관계추출기의 성능을 높이고 지식베이스를 확장한다.

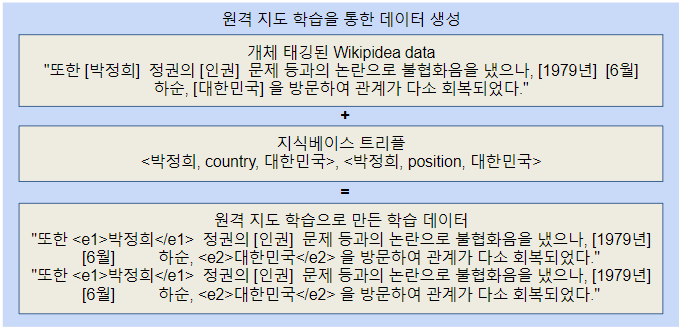

그림 4 지식베이스 트리플들을 시드로 사용하여 개체 연결이 완료된 위키피디아 데이터를 대상으로 원격 지도 학습을 통해 관계추출기의 학습 데이터를 생성하는 예

3.1 학습용 데이터 생성

그림 4와 같이 초기 지식베이스에 저장된 관계들을 가지고 원격 지도 학습을 통해 개체 연결이 완료된 데이터로부터 관계추출기의 학습을 위한 데이터를 생성해낸다. 원격지도학습은, 지식베이스에 두 개체간의 관계가 존재하면, 이 두 개체 쌍을 가지는 문장은 지식베이스에 존재했던 관계가 어느 정도는 표현이 될 것이라는 약한 가정을 가지고, 두 개체가 문장에서 발견되면 그 문장에서 두 개체간의 관계는 지식베이스에 있던 관계라고 관계를 정하여 학습 데이터를 만드는 방식으로, 사람의 노력이 필요 없이 값싸게 많은 데이터를 얻을 수 있는 방식이다.

3.2 관계추출기 학습

관계추출기란, 그림 1과 같이 두 개체를 지정해주면, 문장에서 두 개체간의 의미적 관계가 무엇인지 정해주는 것이다. 이전 단계에서 원격 지도 학습을 통해 생성한 학습용 데이터를 가지고 관계추출기를 학습을 한다. 학습 방식으로는 하나의 문장에서 두 개의 개체가 주어지면 그 두 개체간의 관계가 원격 지도 학습을 통해 정해져 있는데, 관계추출기가 그 관계를 예측 할 수 있도록 지도 학습을 한다. 본 연구에서는 BERT[10] 모델을 이용해 만든 관계추출기를 이용했다.

3.3 관계추출

이 단계에서는 학습된 관계추출기를 이용해 지식베이스에 추가하기 위한 트리플들을 생성하는 단계이다. 개체 연결이 된 문장들이 주어지면 학습된 관계추출기를 이용해 문장에서 개체 간의 관계를 추출해내고, 그 관계를 트리플 형식으로 만든다. 관계추출기의 성능이 좋을수록 더 정확하게 관계들을 정의할 수 있다. 본 연구에서는 한국어 위키피디아 데이터를 대상으로 관계들을 추출하였고, 한 문장에서 두 개의 개체를 선택하는 모든 조합에 대해 관계추출을 진행 하였다.

3.4 지식 검증

이 단계에서는 관계추출기가 추출한 트리플들에 대해 진위여부를 판단하는 단계이다. 관계추출기가 정의한 관계들은 틀린 관계가 많이 섞여있기 때문에 그대로 지식베이스에 넣게 되면 지식베이스의 신뢰도가 많이 낮아지게 된다. 그렇기 때문에 지식 검증을 통해 정확하지 않은 트리플들은 지식베이스에 추가되지 않도록 해야 한다. 이를 위해 본 연구에서는 관계추출기가 제시한 신뢰도가 0.9 이상인 관계들을 대상으로 여러 지식 검증 알고리즘을 거치도록 하였다. 그 중 하나로는 개체 타입 검증인데, 트리플에서 관계의 정의역과 치역에 올 수 있는 타입이 아닌 다른 타입이 오면 거짓인 트리플이라 판단하는 검증이다. 예를 들면 <백남준, country, 지미 카터>와 같이 트리플의 관계가 country인데, 치역으로 개체 타입이 President인 지미 카터가 오게 되면 거짓인 트리플이라 판단한다.

3.5 반복학습 자동화

반복학습은 잘못된 지식이 지식베이스에 쌓이게 되면서 오히려 관계추출기의 성능이 악화될 가능성이 존재한다. 그래서 어떻게 지식 검증을 해야 할지가 중요하고, 각 검증 방법을 통해 실제로 성능이 향상되는지를 확인해 봐야 한다. 그렇기 때문에 반복학습 전체과정을 자동화하여 많은 실험을 적은 노력으로 할 수 있도록 하였다. 각 단계별 모듈들의 핵심 설정들을 하나의 파일에서 관리 할 수 있도록 하고, 각 모듈의 결과값이 다음 모듈의 입력값으로 바로 사용될 수 있도록 맞추었다. 그리고 하나의 실행파일이 각 모듈을 차례로 실행할 수 있도록 하였다.

4. 실험 및 분석

관계추출기의 학습 데이터를 위한 원격 지도 학습 및 지식베이스 확장을 위한 추출 대상 말뭉치는 한국어 위키피디아 20180801 버전의 초록부분을 사용했다. 이 위키피디아 데이터에 개체 연결을 해서 사용했는데, 그를 위한 모듈로는 ELU2018[11] 모델을 사용했다. 시드로 사용된 트리플은 한국어 DBpedia 기반인 KBox를 기반으로 사용했다. 그리고 관계의 종류는 44종류를 사용했고, 관계추출기는 BERT기반의 관계추출기를 사용했다. 관계의 종류는 Kbox의 121개 관계를 44개로 줄인것인데, 관계추출기 성능 평가 결과가 70점 이상인 관계만 선별하였고, 크라우드 소싱으로 데이터를 수집 해 본 결과 원격지도 학습의 노이즈가 절반이 넘지 않는 관계들만 남겼다. 그리고 온톨로지상 상위 관계가 있는 경우에는 그 관계로 치환 하였다.

반복적으로 학습함에 따라 관계추출기의 성능이 좋아지거나 나빠지는 것에는 지식 검증에 많은 영향을 받는다. 그렇기 때문에 본 연구에서는 다양한 방법으로 지식 검증을 하며 반복 학습을 시도해 보았다. 표1 에서는 다양한 방식의 필터링을 통해 어떤 필터링이 어떻게 영향을 주는지 관찰한 것이다. 평가 데이터로는 2993개 문장의 Gold 데이터를 활용했다.

Iter0는 초기 지식베이스에서 44개의 관계에 해당하는 트리플을 가지고 원격 지도 학습을 통해 학습데이터를 만들어 학습한 결과이다. Iter1-1은 Iter0에서 추출한 결과에 대해 점수가 0.9점 이상이면서 관계가 정의하는 정의역과 치역의 타입이 올바른 트리플만 추가하여 실험하였다. 그 결과, 특정한 몇 개의 관계들이 오류 학습데이터를 많이 만들어 내는 경향도 보였다. 그리하여 1-2에서는 오류 데이터가 많이 섞이는 관계인 country와 loaction을 제외해서 등록했다. 그리고 개체 타입이 정의되지 않아 정의역과 치역의 타입 검증이 적용되지 않았던 트리플들도 제외했다. 1-3에서는 1-2에 추가로 두 개체를 선택할 때, SRL[12]의 Predicate-Argument 구조를 사용하여 Predicate에 종속된 두 개체에 대해서만 추출을 진행한 결과이다. 전혀 관계가 없는 두 개체에서 관계를 뽑아야 하는 경우가 많아서, 관계가 있을 가능성이 높은 두 개체를 강제로 선택해 주기 위해서였다. 결과적으로는 Iter1-2와 같은 필터를 적용했을 때가 가장 성능의 향상이 좋았다. Iter2의 경우는 Iter1-2에서 Iter1-1과 같은 조건에 KV모듈[13]을 추가해서 사용한 결과이다. 이와 같은 실험으로 반복학습이 지식베이스의 확장 및 관계추출기의 성능 향상에 기여할 수 있다는 것을 볼 수 있다.

하지만, 표 1에서의 실험에서는 증가된 지식에 많은 오류 지식이 섞여있었기 때문에, 다른 조건으로 표 2와 같은 실험을 하였다. 표 2에서의 실험은 추출 결과에 대해 점수가 0.95점 이상인 트리플을 대상으로, 오류가 많은 몇몇 관계는 추출에서 제외하였고 개체 타입이 정의되지 않은 트리플은 관계에 대한 개체 타입으로 필터링을 할 수 없기 때문에 삭제하였다. 그 뒤 관계에 대한 정의역과 치역의 타입 검증 및 KV모듈을 모두 적용한 실험 결과이다. 그리고 초기 지식베이스에 추가로 226,058개의 트리플을 더 추가 하였고, 학습데이터에는 크라우드소싱을 통해 참으로 판명된 108,333개의 학습 데이터를 붙여서 실험을 하였다. 이 경우에는 관계추출기의 성능은 일정하게 유지되는 경향을 보이고 지식베이스의 확장에서만 성과를 볼 수 있었다.

표 1 반복학습 실험 결과 1

반복	트리플 수	학습데이터 수	Precision	Recall	F1
Iter0	797,046	836,517	0.7597	0.7407	0.7415
Iter1-1	1,040,504	2,820,350	0.7746	0.7738	0.7742
Iter1-2	908,236	1,388,322	0.7821	0.7805	0.7813
Iter1-3	847,664	1,342,533	0.7586	0.7571	0.7579
Iter2	1,528,441	5,353,430	0.7910	0.7905	0.7908

표 2 반복학습 실험 결과 3

반복	트리플 수	학습데이터 수	Precision	Recall	F1
Iter0	1,023,104	1,174,720	0.7616	0.7601	0.7609
Iter1	1,084,880	1,448,783	0.7625	0.7614	0.7620
Iter2	1,192,275	1,647,465	0.7527	0.7514	0.7520
Iter3	1,327,031	1,750,822	0.7581	0.7571	0.7576
Iter4	1,476,859	1,811,147	0.7549	0.7534	0.7542
Iter5	1,638,101	1,854,876	0.7538	0.7531	0.7535
Iter6	1,809,673	1,894,208	0.7555	0.7548	0.7551
Iter7	1,987,720	1,928,767	0.7598	0.7588	0.7598
Iter8	2,172,292	1,959,741	0.7505	0.7497	0.7501
Iter9	2,360,267	1,981,506	0.7565	0.7558	0.7561

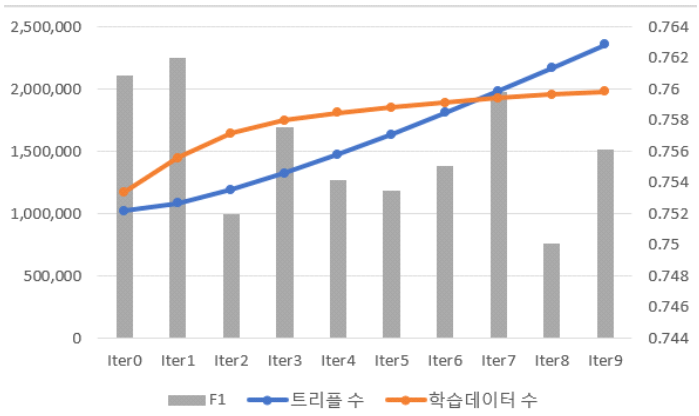

그림 5 표 2의 시각화

5. 결론

본 연구에서는 반복 학습을 이용하여 “지식 추출-> 지식증강 -> 관계추출기 성능 향상 -> 더 양질의 지식 추출” 이라는 선순환을 이루어 내면 관계추출기의 성능을 향상시킬 수 있고, 지식베이스를 확장하는데 도움을 줄 수 있다는 사실을 보였다. 그러나 아직 지식 추출기의 성능 향상 폭이 적고, 오류 지식 데이터가 지식베이스에 저장 되는 경우가 많다. 이를 해결하기 위해서는 지식 검증 알고리즘이 더 확실하게 잘못된 지식을 걸러야 한다. 추후 연구로는 오류 분석을 통하여 어떤 지식 검증 알고리즘을 사용하면 가장 적합하게 위와 같은 선순환을 만들어 낼 수 있는지에 대해 연구하고, 그 경우 어디서 그 결과가 수렴하게 되는지를 알아내는 것이 남아있다.

사사

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2013-0-00109, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

참고문헌

[1] Soren Auer, Christian Bizer, Georgi Kobilarov, JensLehmann, Richard Cyganiak, and Zachary Ives. "Dbpedia: A nucleus for a web of open data." In The semantic web, pages 722 735. - Springer, 2007.

[2] Fabian M S uchanek, Gjergji Kasneci, and GerhardWeikum. "Yago: a core of semantic knowledge." I n Proceedings of the 16th international conference on World Wide Web, pages 697 706. ACM, 2007.

[3] Kurt Bollacker, Colin Evans, Praveen Paritosh, TimSturge, and Jamie Taylor. " Freebase: a collaboratively created graph database for structuring h uman knowledge." In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 1247 1250. AcM, 2008.

[4] Nam, Sangha, Eun-Kyung Kim, Jiho Kim, Yoosung Jung, Kijong Han and Key-Sun Choi. “A Korean Knowledge Extraction System for Enriching a KBox.” COLING (2018).

[5] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. "Distant supervision for relation extraction without labeled data." In Proceedings of the J oint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Vol-ume 2-Volume 2, pages 1003-1011. Association for Computational Linguistics, 2009.

[6] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar,B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner,B. Kisiel, J. Krishnamurthy, N. Lao, K. Mazaitis,T. Mohamed, N. Nakashole, E. Platanios, A. Ritter,M. Samadi, B. Settles, R. Wang, D. Wijaya, A. Gupta,X. Chen, A. Saparov, M. Greaves, and J. Welling,“Never-ending learning,” Commun. ACM, Vol. 61,No. 5, pp. 103-115, Apr. 2018.

[7] W. Jia, D. Dai, X. Xiao, and H. Wu, “ARNOR:Attention regularization based noise reduction fordistant supervision relation classification,” Proceedingsof the 57th Annual Meeting of the Association forComputational Linguistics, pp. 1399-1408, Jul. 2019.

[8] J. Feng, M. Huang, L. Zhao, Y. Yang, and X. Zhu,“Reinforcement learning for relation classification fromnoisy data,”ArXiv, Vol. abs/1808.08013, 2018.
[9] Yi Wu, David Bamman, and Stuart Russell. Adversarial training for relation extraction. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 1778-1783. 2017.
[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova,“Bert: Pre-training of deep bidirectional transformersfor language understanding,”NAACL-HLT, 2018.
[11] Le, Phong; Titov, Ivan. “Improving Entity Linking by Modeling Latent Relations between Mentions”ACL, 2018
[12] Carreras, Xavier, and Lluís Màrquez. "Introduction to the CoNLL-2004 shared task: Semantic role labeling." Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2004) at HLT-NAACL 2004. 2004.
[13] 김지호, 한기종0, 최기선, “KBCNN: CNN을 활용한 지식베이스 완성 모델”, 한글 및 한국어 정보처리 학술대회 논문집 (2018년), 465-469, 2018