EMNLP 2018 참관기

글 | 서가은 annie.mation@kakaocorp.com 〈아이언맨〉의 자비스를 꿈꾸며 현재는 카카오미니에서 자연어 처리를 담당하는 개발자입니다. 자비스까지는 아직 먼길이라 생각하지만 카카오미니가 좀 더 나은 삶에 기여해주길 바라며 열심히 하고 있습니다.

글 | 이다니엘 daniel.lee@kakaocorp.com 카카오에서 자연어 처리와 관련된 일을 하고 있습니다. 데이터를 보고 특징들을 분석하는 것을 좋아합니다. 최근에는 카카오미니와 관련된 일을 하면서 열심히 공부하고 있습니다.

글 | 이동훈 hubert.lee@kakaocorp.com 어쩌다 보니 자연어 처리 분야에서 10년 넘게 일하고 있는 30대 개발자입니다. 형태소 분석이나 개체명 인식과 같은 자연어 처리 기반 기술에 관심이 많습니다. 본격적으로 딥러닝에 입문한 지는 얼마 되지 않아서 모든 것이 새롭고, 신기하게 느껴지고 있습니다. 기계가 사람 말, 특히 한국어를 잘 이해할 수 있는 그날까지 열심히 노력할 것입니다.

개요

EMNLP(Empirical Methods in Natural Language Processing)*1는 자연어 처리에서 경험적 방법론을 다루는 학회로, ACL(Association for Computational Linguistics)과 함께 전산언어학 분야에서는 인지도가 높은 콘퍼런스다. EMNLP 2018은 벨기에 브뤼셀에서 10월 31일부터 11월 4일까지, 5일간 열렸다. 최근 자연어 처리에 대한 높은 관심을 반영하듯 많은 논문과 참석자들로 인해 어느 때보다 관련 내용이 풍성했다. 제출된 2100편의 논문 중 549편이 수락되고, 14개의 워크숍, 6개의 튜토리얼과 3개의 키노트가 개최됐다. 참석자는 약 2500명으로 작년에 비해 두 배 가까이 증가했다고 한다. 본 글은 카카오에서 자연어 처리 기술을 연구하는 이동훈, 이다니엘, 서가은, 세 명이 EMNLP 2018을 다녀온 후 작성한 참관 후기로, 가장 기억에 남았던 세션을 공유하고자 한다.

대화(Dialogue)

처음으로 소개해드릴 주제는 '대화'다. 대화 처리 시스템(dialogue system)은 '자연어로 사람과 대화하는 시스템'이다. 일반적으로 대화 처리 시스템은 언어 이해(natural language understanding), 대화 관리(dialogue management), 언어 생성(natural language generation)으로 구성된다.*2 최근에는 아마존 알렉사(Amazon Alexa), 구글 홈(Google Home), 카카오미니(kakomini) 같은 스마트 스피커(smart speaker)들이 출시되면서 대화 처리 시스템에 대한 연구가 활발히 진행되고 있다. 이번 콘퍼런스에서 다루어진 연구들 중 흥미로운 몇 가지를 다음과 같이 꼽아보았다.

1) 질문의 의도 분류 방법(Supervised Clustering of Questions into Intents for Dialogue System Applications)*3

대화 처리 시스템에서는 사용자의 요청을 처리하기 위해 시스템 설계 시 그 의도(intent)를 미리 분류해놓는다. 본 연구에서는 대화 처리 시스템 설계 시 활용할 수 있는 '질문의 의도를 분류하는 방법'을 제안하고 있다. 짧은 문장의 의미 정보(semantic)를 분류하기 위해 '구조화된 출력(structured output)'을 입력으로 사용하는데, 여기서 구조화된 출력은 질문과 유사도를 그래프 형태로 표현한 것이다. 본 연구를 통해 대화 처리 시스템의 의도 체계를 설계하는 과정이 간단해지면서 동시에 빠르게 진행될 수 있게 되었다. 이에 새로운 대화 처리 시스템의 의도 체계를 설계할 때 본 연구를 활용하면 좋을 것으로 예상한다.

2) 대화를 위한 음성 인식 모델(Session-level Language Modelling for Conversational Speech)*4

음성 대화 처리 시스템(spoken dialogue system)에서 음성 인식 성능을 향상하기 위한 연구다. 대화 음성 인식 시, 항상 새로운 대화가 시작되는 것처럼 음성을 인식하는 것이 아니라 이전 대화 정보를 사용해 음성 인식을 해보자는 것이 이 연구의 핵심 아이디어다. 대화란 매번 새로 시작되는 것이 아니라 이전 발화에 이어지는 것이기 때문이다. 따라서 대화를 주고받는 정보(turn taking)나 발화가 겹치는지(overlap)에 관한 정보를 포함해 이전 대화의 모든 정보를 문맥(context)으로 사용한다([그림 1]). 이 모델은 기존의 대화 음성 인식 모델에 '대화의 본질'을 반영해보려고 한 시도 자체가 의미 있다고 생각된다.

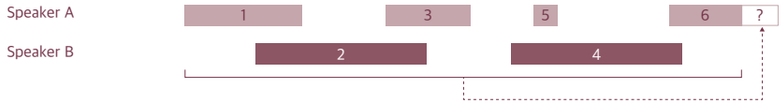
*1 참고 | https://en.wikipedia.org/wiki/Empirical_Methods_in_Natural_Language_Processing

*2 참고 | https://en.wikipedia.org/wiki/Dialogue_system

*3 논문 | Iryna Haponchyk, 'Supervised Clustering of Questions into Intents for Dialog System Applications', EMNLP, 2018.

*4 논문 | Wayne Xiong, 'Session-level Language Modeling for Conversational Speech', EMNLP, 2018.

[그림 1] 이전 발화자의 문맥 정보를 사용하는 대화 음성 인식 모델[*4]



3) 페르소나가 있는 대용량 말뭉치 생성(Training Millions of Personalized Dialogue Agents)[*5]

페르소나(persona)는 성격(character)을 의미한다. 잡담(chit-chat)하는 대화 에이전트(예를 들어 챗봇)를 개발할 때, 페르소나가 있다면 에이전트의 답변이 훨씬 더 사람의 답변처럼 느껴질 것이다. 본 연구는 페르소나가 포함된 말뭉치 생성 방법이다. Reddit[*6]의 데이터를 활용해 대화 말뭉치를 생성하고 몇 가지 조건으로 전처리를 함으로써 미리 정의한 방법을 통해 대화에 포함된 페르소나를 추출한다. 본 연구에서는 페르소나가 포함된 7억 개의 대화 말뭉치를 생성했다고 한다. 참고로 기존 연구에서 생성한 말뭉치는 수십만 개에 불과했다. 본 연구를 한글 데이터에 적용해 페르소나가 포함된 한글 대화 말뭉치를 생성해볼 수 있을 것 같다.

페르소나가 포함된 대화 말뭉치의 예[*5]

· Persona: ["I like sports", "I work a lot"]
· Context: "I love running."
· Response: "Me too! But only on weekends."

[그림 2] 딥러닝이 접목된 챗봇 세션을 듣는 참가자들

개체명 인식(Named Entity Recognition, NER)

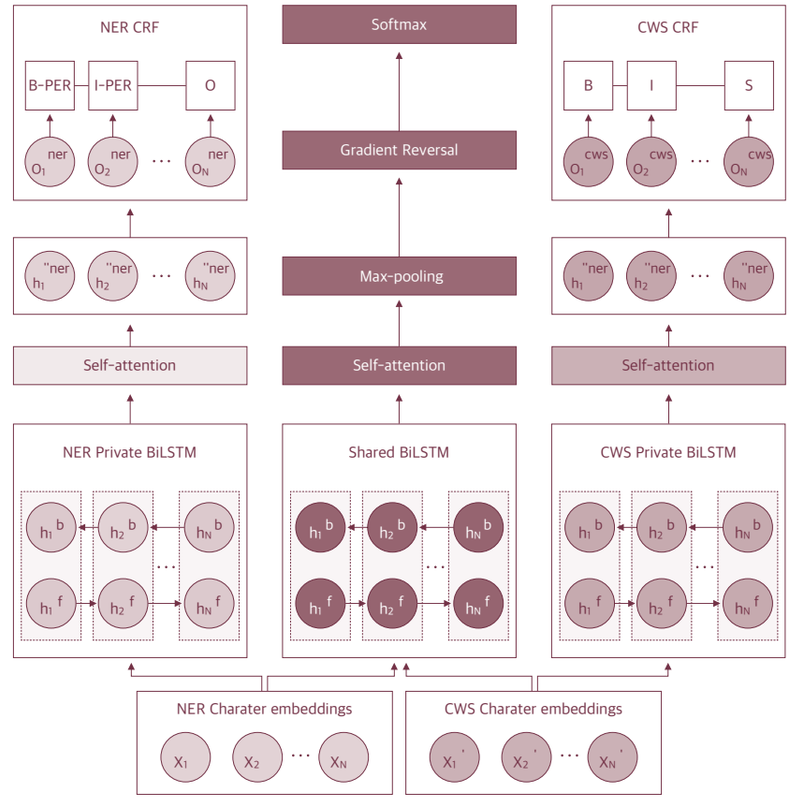
두 번째는 개체명 인식에 관한 연구[*7]다. 이는 오랜 과업이자 어려운 문제이기에 어떤 새로운 방법론이 나왔는지 열심히 살펴보았다. 주로 새로운 모델을 제시하는 논문보다는 적은 양의 학습 데이터를 극복하는 방법과 다중 작업 학습(multi-task learning)을 사용해 2가지 이상의 작업(task)을 한 번에 해결하는 연구가 많았다. 개체명 인식과 단어 경계 분리 작업은 '단어 경계'라는 공통된 피처가 중요하게 사용되는데, 본 연구에서는 이 2가지 작업을 다중 작업 학습으로 만들고 '단어 경계'라는 피처를 공유함으로써 성능을 개선했다.

*5 논문 | Pierre-Emmanuel Mazare, 'Training Millions of Personalized Dialogue Agents', EMNLP, 2018.

*6 참고 | https://www.reddit.com

*7 논문 | Pengfei Cao, 'Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism', EMNLP, 2018.

[그림 3] 개체명 인식과 단어 경계 분리를 동시에 학습하는 모델[*7]



성능 개선을 가능케 한 부분은 바로 [그림 3]의 shared BiLSTM이다. 이는 각 작업에 특화된 피처를 배제하고 공통된 피처만 학습함으로써 두 작업의 성능을 향상시킨다. 특히 단어 경계 분리 작업에 사용되는 코퍼스의 양이 많아서 개체명 경계를 인식하는 데 큰 도움을 줄 수 있다.

한편 개체명 태그 세트가 다른 두 학습 데이터를 활용하는 방법에 관한 연구도 있었다.[*8] 이 연구에서는 다음과 같이 3가지 모델을 제안하고, 그 성능을 비교했다([표 1]).

[표 1] 두 학습 코퍼스를 사용하는 모델 설명 및 성능

모델(Model)	설명(Description)	F-score
CP Single Model	ChemPort 코퍼스 하나만 사용한 경우	82.3%
Single Model	학습 데이터와 개체명 태그 세트를 병합해 처리	83.3%
Multi CRF Model	개체명 태그 세트별로 독립된 CRF-layer로 태깅 후 병합	84.5%
Marginal CRF Model	CRF-layer에서 likelihood 계산 시, 관측된 태그열에 집중	85%

실험에는 CDR(chemical disease relation)와 CP(ChemPort) 2개의 코퍼스가 사용됐다. [표 1]에서 볼 수 있듯이 CP 코퍼스만 사용해서 태깅한 경우보다 CDR 세트를 같이 사용한 케이스의 성능이 적게는 1%, 많게는 2.5% 이상 향상된 것을 확인할 수 있다. 참고로 [표 1]의 F-score는 CP 코퍼스의 protein 레이블 성능을 나타낸다.

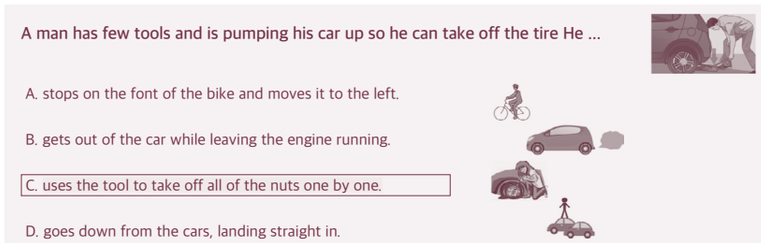
*8 논문 | Nathan Greenberg, 'Marginal Likelihood Training of BiLSTM-CRF for Biomedical Named Entity Recognition from Disjoint Label Sets', EMNLP, 2018.

그 밖에도 AT(adversarial training)를 사용함으로써 개체명 인식과 개체명 연결 작업의 성능을 향상시킨 연구가 있었다.*9 그리고 개체명 연결을 end-to-end 방식으로 접근해 문제를 해결하는 연구도 있었다.*10

자연어 추론(Natural Language Inference, NLI)

세 번째는 자연어 추론이다. 자연어 추론이란 상황을 설명하는 문장이 주어졌을 때 그 상황에 대한 상식적인 추론을 의미한다.*11 예를 들어 [그림 4]와 같이 자동차 타이어를 교환하는 상황이 주어지고, 상황에 맞게 이어질 적절한 문장을 추론하는 것이 바로 자연어 추론이다.*12 이와 관련해서 공개된 데이터 세트(dataset)인 'SWAG(Situations with Adversarial Generations) : A Large-scale Adversarial Dataset for Grounded Commonsense Inference'를 다룬 연구*11가 있었다. 금번 학회에서는 wikiConv*13등 상당히 많은 데이터 세트가 공개됐지만 그중 자연어 추론과 관련된 데이터 세트는 SWAG뿐이었다. 최근 트렌드와는 성격이 다르지만 새로운 연구 주제로 인상 깊었던 논문이었다.

[그림 4] SWAG에서 자연어 추론의 예시*12



SWAG 연구에서 공개한 데이터 세트는 현실에서 생길 수 있는 상황 시나리오를 기반으로 제작됐으며, 비디오 캡션(video caption)에서 추출한 약 11만 개의 객관식 질의다. 이 질의에 대해서는 상황에 대한 문장과 1개의 정답, 그리고 3개의 오답이 존재한다. 오답은 언어 모델(language model)로 과도하게 오류 문장들을 생성하고, Adversarial Filtering*14 방법을 사용해 기계를 헷갈리게 한다. SWAG에서는 이 방법을 포함한 데이터 세트 생성 절차(dataset construction procedure)를 소개하고 있다.*11 이는 복잡하고 어려운 데이터 생성 과정에서 기존의 자연어 추론 데이터 세트가 가지고 있는 코퍼스를 작성해, 잘못된 해석 또는 추론이 작성될 수 있는 등의 문제점*15을 해결했다고 한다.

SWAG는 리더보드를 통해 모델별로 성능을 공개하고 있다.*16 사람의 정확도는 88% 정도이고, 최신 모델인 ESIM+ELMo는 59% 정도라고 한다. 현재 리더보드에서는 BERT(Bidirectional Encoder Representations from Transformers)*17라는 모델이 약 86%의 정확도로 가장 뛰어난 성능을 가지고 있다. 특히 이 연구에서는 벌써부터 사람과의 차이가 거의 없는 모델이 나왔다는 점이 가장 충격적이고 흥미로웠던 부분이었다.

*9 논문 | Giannis Bekoulis, 'Adversarial training for multi-context joint entity and relation extraction', EMNLP, 2018.

*10 논문 | Nikolaos Kolitsas, 'End-to-End Neural Entity Linking', CoNLL, 2018.

*11 논문 | Rowan Zellers, 'Swag: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference', EMNLP, 2018.

*12 참고 | https://drive.google.com/file/d/1vHH9kqufVdTWzFC734VQYd2Gz2DRpIsN/view

*13 참고 | https://aclanthology.coli.uni-saarland.de/papers/D18-1305/d18-1305

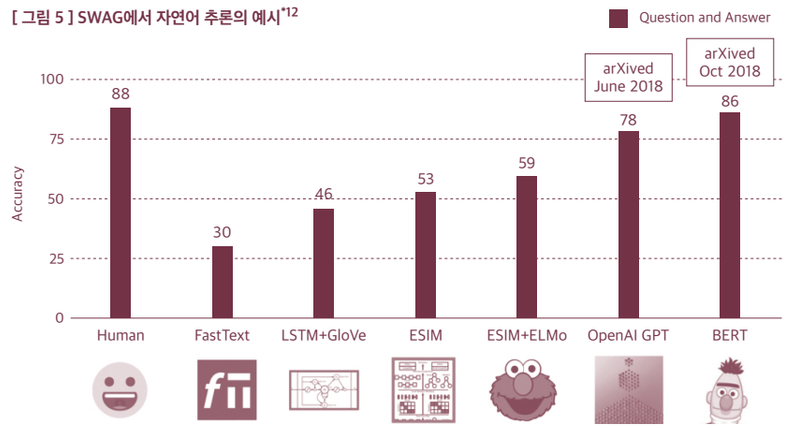
*14 참고 | https://github.com/rowanz/swagaf

*15 논문 | Omer Levy, 'Annotation Artifacts in Natural Language Inference Data', NAACL-HLT, 2018.

*16 참고 | https://leaderboard.allenai.org/swag/submissions/public

*17 논문 | Jacob Devlin, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', 2018.

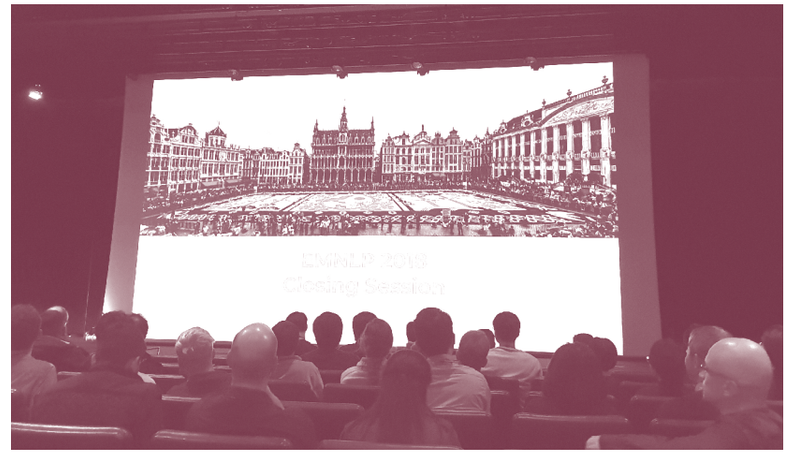
[그림 5] SWAG에서 자연어 추론의 예시*12



마무리

5일이라는 짧지만은 않은 학회를 다녀오면서 각자 느낀 소회를 가볍게 정리하며 이 글을 마무리하려 한다.

[그림 6] EMNLP 2018 클로징 세션



"주관적인 느낌일 수는 있겠지만, 다중 작업 학습, 전이 학습(transfer learning)을 이용한 방법론이 많이 연구된 것 같았습니다. 특히 중국에서 많은 논문이 발표되었는데, 영어권에 있는 풍부한 코퍼스의 전이 학습이 중국어 버전 시스템의 성능 향상에 기여했다는 논문도 나왔습니다. 다중 작업 학습에 대해서는 회의적인 느낌도 있었지만, 우수 논문(best paper)에 선정된 논문들을 포함해 많은 곳에서 기술 개발과 관련한 성과들이 보인다는 내용들이 있으니, 서비스 적용에 시도를 해봐도 좋지 않을까라는 생각을 해봤습니다." -이동훈

"평소 주어진 업무에 몰두하다 보면 하나의 문제를 고민하느라 다른 문제들을 고민하지 못하는 경우가 많습니다. 서비스와 밀접한 업무를 진행할 때는 다른 부차적인 것들로 인해 더더욱

그러합니다. 이번 EMNLP 학회에 참석하면서 다양한 주제의 연구 결과를 보고 느끼며 평소에 아쉬웠던 부분들을 해소할 수 있었습니다. 특히 잘 준비된 발표들로 논문을 읽는 것보다 더욱 효과적으로 연구 결과들을 이해할 수 있었던 것 같습니다." -이다니엘

"이번 EMNLP에서는 adversarial methods를 적용한 논문들을 적지 않게 찾아볼 수 있었고, ELMo와 같은 문맥 임베딩(contextual embeddings)을 사용하는 논문이 많았습니다. 더불어 다양한 데이터 세트도 많이 소개됐고, QA와 dialogue 등 다양한 주제의 논문들도 많이 발표되었습니다. 점점 더 풍성해지는 자연어 처리 연구들을 보면서, 장기적으로 해야 할 일도 많고 더 많이 공부해야겠다는 자극을 받을 수 있었습니다. 최신 연구 동향들을 살펴보면서 더 똑똑한 자연어 처리 기술을 만들어야겠다는 생각도 했습니다." -서가은

〈EMNLP 2018 참관기〉 브런치로 연결되는 QR 코드입니다.