채팅체-문어체 스타일 변환 병렬 코퍼스 자동 구축

최민주O, 이창기, 황정인, 노형종

강원대학교 빅데이터메디컬융합학과, NCSOFT
mjchoi0831@gmail.com, leeck@kangwon.ac.kr, {jihwang, nohhj0209}@ncsoft.com

Automatic Construction of Chat-Written Style Conversion Parallel Corpus

Minjoo ChoiO, Changki Lee, Jeongin Hwang, Hyungjong Noh
Dept. of Medical Bigdata Convergence, Kangwon National University
NCSOFT

요 약

인터넷 채팅체로 쓰여진 문장은 문어체 문장과 달리 신조어 및 축약어가 쓰이며 문체 또한 일반적인 문어체 또는 구어체와 상이하다. 따라서 인터넷 채팅체를 기존 문어체 기반 자연어처리 시스템에서 이용하기 위해서는 채팅체-문어체 스타일 변환 기술이 필요하며, 이를 위해서 구어체 - 문어체로 이루어진 병렬 코퍼스를 구축할 필요가 있다. 본 논문에서는 채팅체 문장을 문어체로 변환한 문장 쌍 병렬 코퍼스를 Round-Trip Translation 기법을 이용하여 자동으로 구축하고, 자동으로 구축된 병렬 코퍼스 중에 부정확한 문장 쌍을 자동으로 필터링하는 방법을 제안한다. 또한 구축된 병렬 코퍼스를 검증하기 위해 구축된 병렬 코퍼스를 이용하여 자동으로 채팅체-문어체 변환 사전을 구축하였다.

주제어: 데이터 증강, Data Augmentation, Round-Trip Translation, 기계번역, NMT

1. 서론

게임 채팅 및 인터넷 게시판에서 사용하는 문장에는 신조어 및 축약어가 쓰일 뿐만 아니라 문체도 일반 문어체나 구어체와 상이하기 때문에 보편적으로 자연어처리 시스템에 이용되는 문어체 문장과 다르다. 예를 들어 채팅체 문장들에 단순히 기존의 자연어처리 시스템을 적용할 경우 오류가 발생할 수 있다. 이를 보완하기 위해 채팅체로 구성된 학습데이터를 구축하여 자연어처리 시스템에 추가로 학습시키면 성능이 개선될 수 있으나, 학습데이터를 구축하기 위해서는 상당한 비용과 시간이 소요될 수 있다.

그러므로 채팅체를 문어체로 자동으로 변환할 수 있다면 수작업으로 말뭉치를 구축하여 자연어처리 시스템에 학습시키는 경우에 비해 상대적으로 적은 시간과 비용을 들여 자연어처리 시스템의 성능을 개선하는 효과를 기대할 수 있다.

텍스트 스타일 변환(text style transfer)은 입력 문장의 내용은 유지하면서 문장의 스타일을 변환하는 작업으로, 텍스트 스타일 변환을 위한 병렬 데이터가 충분하다면 기존의 기계 번역 모델을 이용할 수 있지만, 텍스트 스타일 변환 병렬 데이터가 부족하기 때문에 많은 연구[1]에서 비지도 접근 방식에 초점을 맞추고 있다.

본 논문에서는 채팅체를 문어체의 스타일로 변환하여 채팅체-문어체 문장 쌍으로 구성된 병렬 코퍼스를 자동으로 구축하기 위해 다음과 같은 방법을 제안한다.

먼저 채팅체 문장을 수집한 후 Round-Trip Translation (텍스트를 다른 언어로 번역하고, 그 결과를 원래의 언어로 다시 번역하는 과정)을 이용해 채팅체-문어체 병렬 코퍼스를 자동으로 구축하고, 초기 병렬 코퍼스로부터 부정확한 문장 쌍을 자동으로 필터링하였다. 또한 구축된 병렬 코퍼스를 검증하기 위해 구축된 병렬 코퍼스로부터 채팅체-문어체 간 변환 사전을 자동으로 구축하였다.

2. 관련 연구

텍스트 스타일 변환을 위한 병렬 데이터가 부족하기 때문에 많은 연구[1]에서 비지도 접근 방식에 초점을 맞추고 있다. Sequence-to-sequence 모델 기반 텍스트 스타일 변환의 최근 연구[2] 에서는 학습을 위한 병렬 데이터가 부족하여 좋은 성능을 보이지 못하고 있다.

선행된 스타일 변환(style transfer)을 위한 병렬 데이터 확장 연구[3]에서는 뉴스 기사와 같은 문어체로 훈련된 기계번역 모델을 이용하여 구어체와 같은 비형식적인 문장에 Round-Trip Translation을 적용하면 문어체 형식 문장으로 변환되는 경향이 있다는 현상을 이용하여 구어체-문어체 스타일 변환을 위한 병렬 코퍼스를 자동으로 구축하였다.

본 논문에서는 Round-Trip Translation 기법을 이용하여 게임 채팅체 문장을 문어체로 변환하여 초기 병렬 코퍼스를 구축하고, 이로부터 부정확한 문장 쌍을 자동으로 필터링하는 방법을 제안한다.

3. 텍스트 스타일 변환 병렬 코퍼스 자동 구축

3.1 온라인 게임 채팅 코퍼스 수집

온라인 게임 채팅에서 쓰이는 문장을 수집하기 위해 대표적인 게임 커뮤니티인 playNC[1] 와 Inven[2] 사이트의

이 논문은 엔씨소프트의 "인터넷 채팅체에 적응하는 말뭉치 자동 확장 연구" 과제의 지원을 받아 수행된 연구임

[1] https://www.plaync.com/
[2] http://www.inven.co.kr

게시글을 크롤링하였다. 크롤링 대상 게임은 총 6가지로, Aion, Blade&Soul, Lineage, Lineage 2, Lineage M, Lineage 2M 게임을 대상으로 총 4,418,602 단어, 21,370,744 자로 이루어진 716,752 개의 문장을 수집하였다.

3.2 Round-Trip Translation을 이용한 텍스트 스타일 변환 병렬 코퍼스 자동 구축

초기 병렬 코퍼스를 구축하기 위해 수집한 구어체 게임 채팅 문장에 Round-trip translation을 적용하여 채팅체-문어체 문장 쌍을 얻는다.

원문, 즉 한국어로 작성된 게 임 채팅 문장을 Google Translate API를 이용하여 영어로 번역한 후, 다시 한국어로 번역한다. 이 과정을 통해 채팅체 - 문어체로 변환된 병렬 문장 쌍을 생성한다.

표 1 Round-Trip Translation 적용 예시

게임 데이터 (채팅체)	보통 사전예약이면 클로즈베타인건가요
한 → 영 번역	Is it usually closed beta if it is pre-booked?
영 → 한 번역 (문어체)	사전 예약 된 경우 일반적으로 클로즈 베타입니까?
게임 데이터 (채팅체)	모바일게임 많이해본사람들 알겠지만 6 개월이면 꽤나 긴 시간임.
한 → 영 번역	As people who have played a lot of mobile games know, 6 months is quite a long time.
영 → 한 번역 (문어체)	모바일 게임을 많이 해본 사람들이 알다시피 6 개월은 꽤 긴 시간입니다.

4. 텍스트 스타일 변환 병렬 코퍼스 자동 필터링

3장에서 구축한 초기 텍스트 스타일 변환 병렬 코퍼스로부터 정확한 문장 쌍을 자동으로 필터링하기 위해 각 필터링 기법을 Pipeline 형태로 순차적으로 적용한다. 필터링의 최종 결과물은 초기 코퍼스의 약 34%에 해당하는 246,840 개의 문장 쌍이 남는다.

표 2 순차적 필터링 적용에 따른 문장 쌍 개수

필터링 방법	필터링 후 문장 쌍 개수
문장 길이	550,127
BLEU score	522,758
KorNLI Classifier (BERT)	274,371
Bert-score	246,840

4.1 문장 길이를 이용한 필터링

긴 문장의 경우 정확도가 낮으므로 입력 문장 및 스타일 변환 문장의 길이를 100 이하로 제한한다. 또한 번역이 문장의 앞부분만 되는 경우를 제거하기 위해 입력 문장 길이의 90% 보다 번역된 스타일 변환 문장 길이가 짧은 경우를 제거한다.

표 3 문장 길이를 이용한 필터링 예시

원문	특별한 성장물약 질문드려봅니다
스타일 변환	특수 성장 물약

4.2 BLEU score 를 이용한 필터링

'원문-스타일 변환' 문장 쌍의 BLEU score[4]를 측정하여 원문과 번역 결과가 완전 일치하는 경우와 (BLEU = 100) 원문과 번역 결과가 상이하게 다른 경우를 제거한다 (BLEU = 0).

표 4 BLEU score=0 인 경우를 이용한 필터링 예시

원문	막피모집중
스타일 변환	신병 모집
원문	좀 랜덤으로 뜨게라도 해놓지
스타일 변환	나는 그것을 조금 무작위로 만들었습니다.

4.4 KorNLI 를 이용한 필터링

KorNLI[5]는 한국어 Natural language Inference(NLI) 벤치마크 데이터 세트로 Entailment(함의), Neutral(중립), 또는 Contradiction(반대)에 해당하는 문장 쌍으로 구성되어 있다.

KorNLI 를 학습한 BERT Classifier 를 이용해 자동 구축된 텍스트 스타일 변환 병렬 코퍼스로부터 문장 쌍의 관계가 Neutral 또는 Contradiction 인 쌍(즉, 원문-스타일 변환 문장 간 의미가 같지 않은 쌍)을 제거한다. 또한 문장 쌍의 관계가 Entailment 로 의미가 같다고 분류된 경우에도 BERT Classifier 의 분류 점수가 0.8 보다 낮은 경우 정확도가 낮으므로 제거한다.

표 5 KorNLI 를 이용한 필터링 예시

Entailment (score ≤ 0.8)	원문	먼말인지알지?
	스타일 변환	멀리 있다는 것을 알고 있습니까?
	원문	간만에 와봤는데 뭔
	스타일 변환	나는 한동안 거기에 있었지만 뭐
Neutral	원문	집나가면 개 고 생
	스타일 변환	집에 가면 개가 고통스러워
	원문	잣됐따!!!!!!!!
	스타일 변환	맛있어요 !!!!!!!!
Contradiction	원문	베터리 광탈 하니?
	스타일 변환	당신은 배터리 버프입니까
	원문	안치고 같이 놀던데
	스타일 변환	같이 놀지 않았어

4.5 BERT-score 를 이용한 필터링

BERT-score[6]는 두 문장의 유사도를 비교하기 위한 자동 평가 지표이다. 사전 학습된 Contextual embedding BERT 를 활용하였으며 코사인 유사도를 이용하여 병렬 문장 쌍 간의 단어를 매치하는 방법으로 precision, recall, F1 점수를 각각 계산하여 두 문장의 유사도를 측정한다. 두 문장 간의 유사도 점수가 0.9 보다 낮은 문장 쌍을 제거한다.

표 6 BERT-score를 이용한 필터링 예시

원문	이게 오토들 정리 한거냐
스타일 변환	오토가 조직 되었나요?
원문	운영진짜 머같이하네
스타일 변환	작동은 실제로 동일합니다

5. 채팅체-문어체 변환 사전 자동 구축

3장 및 4장의 과정을 거쳐 만들어진 스타일 변환 병렬 코퍼스를 검증하기 위해 구축된 병렬 코퍼스로부터 채팅체-문어체 간 변환 사전을 자동으로 구축한다.

5.1 단어 - 단어 변환 사전 구축

구축된 스타일 변환 병렬 코퍼스에 Fast-Align[7]을 적용하여 단어 정렬(Word-Alignment)를 수행하였다. 이어서 단어 정렬 결과로 얻은 단어 간 변환 빈도를 이용하여 어떠한 단어가 특정 단어로 변환될 확률까지 포함하는 채팅체-문어체 단어 변환 사전을 자동으로 구축하였다. 표 7 은 구축된 스타일 변환 병렬 코퍼스의 예제이고, 표 8 은 자동 구축된 단어-단어 변환 사전의 예시이다.

표 7 구축된 스타일 변환 병렬 코퍼스 예제

원문	스타일 변환
일반 **유저**들은 사냥 할 수 없는	일반 **사용자**가 사냥 할 수없는
무과금자동메크로 캐릭	무료 자동 매크로 캐릭터
스펙별 애매한 상황이 많기에	**사양**마다 애매한 상황이 많기 때문에
비정상적인 **시세**에	비정상적인 시장 **가격**으로

표 8 단어 - 단어 사전 예시

원문	번역문장	변환확률	원문	번역문장	변환확률
유저	**사용자**	76.96	**스킬**	**기술**	39.76
캐릭	**캐릭터**	65.64	**시세**	**가격**	38.86
스펙	**사양**	55.56	**리니지**	**Lineage**	37.89
혈맹	**클랜**	55.22	는데	지만	37.88
어요	습니다	52.17	**전설**	**레전드**	37.29
나요	습니까	45.63	**컨텐츠**	**콘텐츠**	37.11
컬렉	컬렉션	45.61	근데	하지만	35.48
네요	습니다	43.53	**다이아**	**다이아몬드**	35.06
엔씨	**NC**	43.18	한테	에게	33.33

5.2 구 - 구 변환 사전 구축

5.1 에서 얻은 단어-단어 정렬 결과를 Phrase-Align, 즉 구-구 형태로 확장한다. 단어 정렬 결과에 Phrase Extraction[3] 을 적용하여 구-구 정렬을 구하고, 전체 변환 빈도를 이용하여 어떠한 구가 특정 구로 변환될 확률까지 포함하는 채팅체-문어체 구 단위 변환 사전을 자동으로 구축하였다. 표 9 는 변환 확률이 모두 100% 인 예이다.

표 9 구 - 구 변환 사전 예시

원문	번역문장	원문	번역문장
유저	사용자	이 다 .	입 니다 .
스펙	사양	**다크셋**	**다크 세트**
캐릭	캐릭터	한테	에게
세 요 .	십 시오 .	다이아	다이아몬드
엔씨	NC	**업뎃**	**업데이트**
카톡	**카카오 톡**	시세	가격
팔 아요	**판매**	**악몽셋**	**악몽 세트**
는 거	는 것	있 나요	있 습니까
근데	하지만	하 나요	니까

[3] https://stackoverflow.com/questions/25109001/phrase-extraction-algorithm-for-statistical-machine-translation

6. 결론

본 논문에서는 채팅체를 문어체로 변환하기 위해, 수집한 채팅체 문장에 Round-Trip Translation을 적용해 문어체로 변환하여 채팅체-문어체 문장 쌍을 구축하고, 부정확한 문장 쌍을 자동으로 필터링하는 방법을 제안하였다. 또한 완성된 병렬 코퍼스를 이용하여 단어 간, 구 간의 스타일 변환 사전을 자동으로 구축하였다.

향후 연구로는, 본 논문에서 구축한 채팅체-문어체 스타일 변환 병렬 데이터와 사전 학습(pre-training) 등을 이용하여 기계 번역 모델 기반의 텍스트 스타일 변환 모델을 개발할 예정이다.

참고문헌

[1] Luo, Fuli, et al. "A dual reinforcement learning framework for unsupervised text style transfer." arXiv preprint arXiv:1905.10060 (2019).

[2] Jhamtani, Harsh, et al. "Shakespearizing modern language using copy-enriched sequence-to-sequence models." arXiv preprint arXiv:1707.01161 (2017).

[3] Zhang, Yi, Tao Ge, and Xu Sun. "Parallel data augmentation for formality style transfer." arXiv preprint arXiv:2005.07522, 2020.

[4] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation. " Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.

[5] HAM, Jiyeon, et al. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. arXiv preprint arXiv:2004.03289, 2020.

[6] ZHANG, Tianyi, et al. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675, 2019.

[7] Dyer, Chris, Victor Chahuneau, and Noah A. Smith. "A simple, fast, and effective reparameterization of ibm model 2." Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013.