(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2025-0098326
(43) 공개일자 2025년07월01일

(51) 국제특허분류(Int. Cl.)
G06F 40/56 (2020.01) *G06F 40/268* (2020.01)
G06N 3/0455 (2023.01) *G06N 3/0985* (2023.01)
(52) CPC특허분류
G06F 40/56 (2020.01)
G06F 40/268 (2020.01)
(21) 출원번호 10-2023-0189496
(22) 출원일자 2023년12월22일
심사청구일자 2023년12월22일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스 애기능생활관 311호 NLP&AI연구실
서재형
서울특별시 동대문구 안암로26길 9, 305호
(74) 대리인
김등용

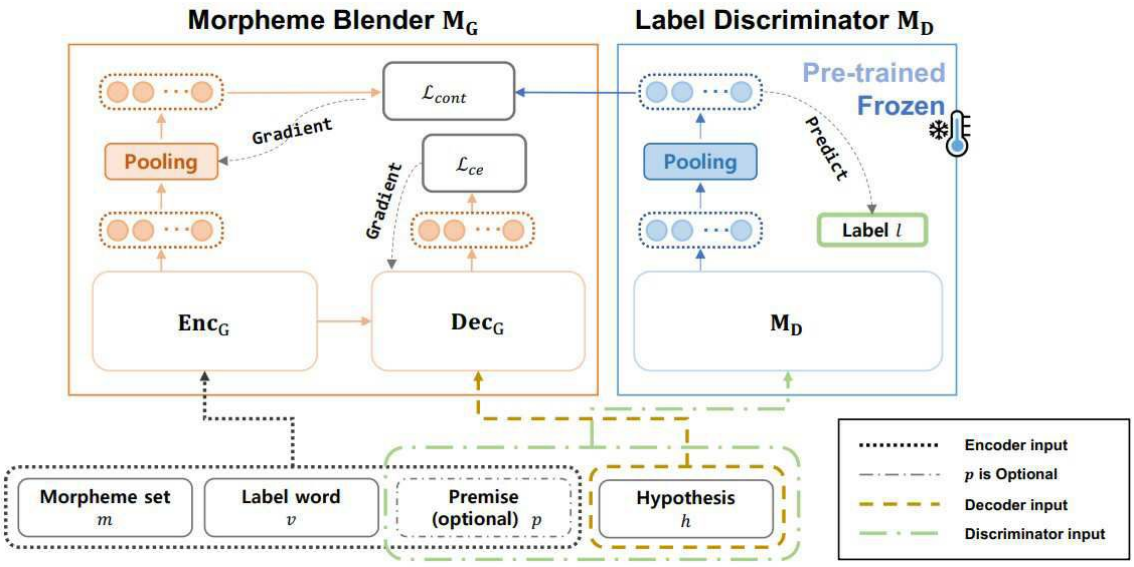
전체 청구항 수 : 총 6 항

(54) 발명의 명칭 **대조적 학습 및 형태소 조합 기반의 한국어 데이터 증강 장치 및 방법**

(57) 요 약

대조적 학습(Contrastive Learning) 및 형태소 조합 기반의 한국어 데이터 증강 장치 및 방법이 개시된다. 상기 한국어 데이터 증강 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 제1 신경망 모델을 학습시켜, 입력되는 문장에 대한 레이블을 추론하는 레이블 판별 모델을 생성하는 단계, 및 상기 레이블 판별 모델의 출력을 이용하여, 제2 신경망 모델을 학습시켜, 입력되는 문장에 포함되는 형태소 집합과 입력되는 문장에 대한 레이블에 대한 자연어 표현인 레이블 워드(Label work)를 입력받아, 형태소 집합을 포함된 형태소를 모두 포함하는 합성 문장을 생성하는 형태소 조합 모델을 생성하는 단계를 포함한다.

대 표 도 - 도2



(52) CPC특허분류
G06N 3/0455 (2023.01)
G06N 3/0985 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1711193821
과제번호 2020-0-00368-004
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 차세대인공지능핵심원천기술개발
연구과제명 뉴럴-심볼릭(neural-symbolic) 모델의 지식 학습 및 추론 기술 개발
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2023.01.01 ~ 2023.12.31
이 발명을 지원한 국가연구개발사업
과제고유번호 1345365402
과제번호 2021R1A6A1A03045425
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축
연구과제명 Human-inspired AI 연구소
기 여 율 1/2
과제수행기관명 고려대학교
연구기간 2023.03.01 ~ 2024.02.29

명 세 서

청구범위

청구항 1

적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되는 한국어 데이터 증강 방법에 있어서,

제1 신경망 모델을 학습시켜, 입력되는 문장에 대한 레이블을 추론하는 레이블 판별 모델을 생성하는 단계; 및

상기 레이블 판별 모델의 출력을 이용하여, 제2 신경망 모델을 학습시켜, 입력되는 문장에 포함되는 형태소 집합과 입력되는 문장에 대한 레이블에 대한 자연어 표현인 레이블 워드(Label work)를 입력받아, 형태소 집합을 포함된 형태소를 모두 포함하는 합성 문장을 생성하는 형태소 조합 모델을 생성하는 단계를 포함하는,

한국어 데이터 증강 방법.

청구항 2

제1항에 있어서,

상기 제1 신경망 모델은 디코더 구조를 갖고,

상기 제2 신경망 모델은 인코더-디코더 구조를 갖는,

한국어 데이터 증강 방법.

청구항 3

제2항에 있어서,

상기 형태소 조합 모델은 제1 손실 함수(L_{ce})를 이용하여 학습되고,

상기 제1 손실 함수는 $$\mathcal{L}_{ce} = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{N}\sum_{j}\log P_{\mathrm{M_G}}(h_j^i \mid h_{<j}^i; seq^i)$$ 이고,

상기 D는 학습 데이터셋, 상기 seq^i는 i번째 문장에 포함된 형태소 집합과 i번째 문장의 레이블 워드를 연결한 시퀀스를, 상기 h^i는 i번째 문장을, 상기 M_G는 상기 형태소 조합 모델을, 상기 N은 상기 D에 포함된 문장의 개수를 나타내는,

한국어 데이터 증강 방법.

청구항 4

제3항에 있어서,

상기 형태소 조합 모델은 제2 손실 함수(L_{cont})를 더 이용하여 학습되고,

상기 제2 손실 함수는 $$\mathcal{L}_{cont} = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{N}\mathcal{L}_{cont}^{i}$$ 이고,

상기 $$\mathcal{L}_{cont}^{i} = -\log\frac{\exp(\mathrm{sim}^i(\mathrm{inp}^i)/\tau)}{\sum_{(\mathrm{inp}^j, l^j)\in C^i}\exp(\mathrm{sim}^i(\mathrm{inp}^j)/\tau)}$$ 이고,

상기 C_i는 상기 D에 포함되는 대조 샘플 집합을, 상기 inp^j는 상기 D 내의 j번째 문장을, 상기 τ는 온도 파라미터(temperature parameter)를 나타내고,

상기 $\mathrm{sim}^{i}(\mathrm{inp}) = \frac{\mathbf{r}_{\mathrm{G}}^{i} \cdot \mathbf{M}_{\mathrm{D}}(\mathrm{inp}))}{\|\mathbf{r}_{\mathrm{G}}^{i}\| \|\mathbf{M}_{\mathrm{D}}(\mathrm{inp}))\|}$ 이고,

상기 $\mathbf{r}_{\mathrm{G}}^{i} = \mathbf{SoftMax}(\mathbf{Enc}_{\mathrm{G}}(seq^{i}) \cdot \mathrm{W}_{\mathrm{G}})$ 이고,

상기 M_{D}는 상기 레이블 판별 모델을, 상기 $\mathrm{Enc}_{\mathrm{G}}$는 상기 형태조 조합 모델의 인코더를, 상기 W_{G}는 상기 MD 내에서 인코딩된 표현을 레이블 분류 확률로 매핑하는 선형 풀링 레이어(Linear pooling layer)를 나타내는,

한국어 데이터 증강 방법.

청구항 5

제4항에 있어서,

타겟 데이터에 대한 형태소 분석을 수행하여, 상기 타겟 데이터를 이루는 타겟 문장에 포함된 형태소 집합을 추출하는 단계; 및

상기 타겟 문장에 포함된 형태소 집합과 상기 타겟 문장의 레이블에 대한 레이블 워드를 상기 형태소 조합 모델에 입력하여, 증강 데이터를 생성하는 단계를 더 포함하는,

한국어 데이터 증강 방법.

청구항 6

제5항에 있어서,

상기 증강 데이터는 상기 타겟 데이터에 포함된 형태소를 모두 포함하는,

한국어 데이터 증강 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 자연어 처리 연구 분야 중에서 자연어 생성에 관한 것으로, 특히 딥러닝 기반의 언어 모델을 활용하여 문장 단위의 데이터를 증강하여 데이터 부족 문제를 극복하고 새로운 데이터 제작에 대한 비용을 절감하며 자연어 이해 태스크를 수행하는 모델의 강건성(robustness)을 높이는 데 활용 가능한 기술에 관한 것이다.

배 경 기 술

[0002] 교착어(agglutinative language)로써, 한국어는 다양한 기능 형태소(functional morphemes)를 포함한다. 한국어 NLP(Natural Language Processing, 자연어 처리) 분야에서, 딥러닝 기반의 연구는 형태소 분리(morpheme segmentation)를 통해 언어 구성 효율성(linguistic construction efficiency)을 향상시키고자 한다. 이전의 연구들은, 모델 성능을 향상시키기 위해, 형태론적 분석(morphological analysis)을 탐구하였다. 한국어는 실질 형태소(lexical morpheme)와 기능 형태소 간의 조합 패턴에 따라 변형이 발생하는 독특한 도전 과제를 제시한다. 따라서, 조합 규칙을 따라 형태학적 변형(morphological alterations)을 통합함으로서 주어진 실질 형태소로부터 다양한 언어 형태(linguistic forms)를 생성할 수 있다.

[0003] 이런 의미에서, 한국어 데이터셋은 실질 형태소와 기능 형태소 간 상호작용을 규제하는 정교한 규칙을 포착하는 데 제약이 있다. 결과적으로, 이러한 데이터셋으로부터 도출된 표현(representations)은 잠재적인 언어 형태의 일부분만을 포함하고 있다. 예컨대, 도 1은 다양한 끝맺음 형태(ending forms, 어미 형태)를 사용함으로써 "먹다(eat)"의 개념을 표현하는 여러 접근법을 보여준다. 또한, 이전의 데이터 증강 접근법은 한국어의 형태학적 특성을 고려한 합성 데이터를 생성하기 위해 드물게 설계되었기 때문에, 데이터 증강에서의 효과가 제한적이다. 이러한 문제를 해결하기 위해서, 명시적으로 형태학적 다양성(morphological diversity)을 고려하는 데이터 증강 방법을 개발하는 것이 중요하다. 이러한 속성을 증강 프로세스에 통합함으로써 모델의 강건성에 기여하고 성능 향상을 촉진할 수 있다.

[0004] 본 발명에서는, 실질 형태소와 기능 형태소의 조합을 이용하여 새로운 합성 데이터(synthetic data)를 생성하기 위한 데이터 증강 방법(CHEF로 명명될 수 있음)을 제안한다. 제안 기법(CHEF)은 형태소 조합기(morpheme

blender, 형태소 조합부나 형태소 조합 모듈 등으로 명명될 수 있음))와 레이블 판별기(label discriminator, 레이블 판별부나 레이블 판별 모듈로 명명될 수 있음)로 구성된다. 이 모듈들(형태소 조합기와 레이블 판별기)은 합성 데이터가 새로운 형태소 조합을 통합하면서도 레이블 일관성(label consistency)을 유지하도록 한다.

[0005] 형태소 조합기는 학습 데이터셋으로부터 추출된 한국어 실질 형태소를 재료로 이용한다. 이러한 실질 형태소는, 파생 및 활용 규칙(derivational and inflectional rule)의 지식을 활용하여, 사전학습된 생성 모델(generative model)로부터 획득된 기능 형태소와 함께 새로운 합성 데이터를 생성하기 위해 이용된다.

[0006] 레이블 판별기는 저장된 실질 형태소를 이용하여 합성 데이터 생성 프로세스를 제어한다. 주요 목적은 의미의 중요한 변화를 방지하고 오리지널 레이블과의 일치를 보장하는 것이다. 형태소 조합기의 티처 포싱(teacher forcing)은 실질 형태소와 기능 형태소의 조합으로 완성된 합성 데이터에 대한 전역 정보(global information)를 통합한다. 이는 레이블 판별기와의 대조적 학습(contrastive learning)을 통해 달성될 수 있다. 도 1에 도시된 바와 같이, CHEF 증강은 오리지널 문장의 실질 형태소 집합에 다양한 기능 형태소를 부착함으로써, 레이블 정보를 유지하면서 텍스트 데이터를 풍부하게 한다.

[0007] 도 1을 참조하면, 한국어 동사(verbs) "먹다", "먹는다", "먹어요", 및 "먹습니다"는 "eat"과 동일한 의미를 갖는다. 형상 내의 색상 변화는 기능 형태소와의 결합으로 인한 형태학적 변화(morphological changes)를, 레시피 N은 다양한 문장을 구성할 수 있는 형태소 집합을, "+1"은 레이블 정보를 나타낸다.

선행기술문헌

비특허문헌

[0008] (비특허문헌 0001) Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. Commongen: A constrained text generation challenge for generative commonsense reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1823-1840.

(비특허문헌 0002) Jaehyung Seo, Seounghoon Lee, Chanjun Park, Yoonna Jang, Hyeonseok Moon, Sugyeong Eo, Seonmin Koo, and Heui-Seok Lim. 2022. A dog is passing over the jet? a text-generation dataset for korean commonsense reasoning and evaluation. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 2233-2249.

(비특허문헌 0003) Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

(비특허문헌 0004) Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871-7880, Online. Association for Computational Linguistics.

(비특허문헌 0005) Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021b. Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.

발명의 내용

해결하려는 과제

[0009] 본 발명이 이루고자 하는 기술적인 과제는 대조적 학습 및 형태소 조합 기반의 한국어 데이터 증강 장치 및 방법을 제공하는 것이다.

과제의 해결 수단

[0010] 본 발명의 실시예들에 의하면, 대조적 학습 및 형태소 조합 기반의 한국어 데이터 증강 장치 및 방법이 제공된다.

[0011] 상기 한국어 데이터 증강 방법은 적어도 프로세서(Processor)를 포함하는 컴퓨팅 장치에 의해 수행되고, 제1 신경망 모델을 학습시켜, 입력되는 문장에 대한 레이블을 추론하는 레이블 판별 모델을 생성하는 단계, 및 상기 레이블 판별 모델의 출력을 이용하여, 제2 신경망 모델을 학습시켜, 입력되는 문장에 포함되는 형태소 집합과 입력되는 문장에 대한 레이블에 대한 자연어 표현인 레이블 워드(Label work)를 입력받아, 형태소 집합을 포함된 형태소를 모두 포함하는 합성 문장을 생성하는 형태소 조합 모델을 생성하는 단계를 포함한다.

발명의 효과

[0012] 본 발명은 대조적 학습 및 형태소 조합을 통해서 한국어 합성 데이터를 제작할 수 있는 언어 모델과 이를 이용한 한국어 데이터 증강 장치 및 방법에 관한 것이다. 초거대언어모델 시대로 접어들면서 데이터 중심의 연구가 핵심으로 부상하기 시작했으며, 이에 따라서 최대한 많은 데이터를 확보하고 사회적 비용도 나날이 높아지고 있다.

[0013] 본 발명에서 시도하는 한국어 데이터 증강은 기존에 수집한 데이터에 대한 품질 하락은 최소화하면서, 교착어의 특성을 반영하여 문장의 다양성을 크게 높이는 방법이다. 또한, 원본 데이터의 크기를 별다른 제약사항 없이 확장하면서, 데이터 수집이 불가능하거나 제약사항이 뚜렷한 분야에 대해서 자연어처리 모델의 성능을 개선하고 강건성을 높일 수 있다. 구체적으로 다음과 같은 효과가 있다.

[0014] (1) 데이터 증강을 통해 전체 훈련 데이터를 사용하는 실험인 풀샷(Full-shot) 설정에서의 성능 향상에 어려움이 있음에도 불구하고, 제안하는 발명은 효과를 입증하였다

[0015] (2) 레이블 유지 장치와 형태소 조합기 사이의 대조적 학습을 활용하는 것이 레이블 일관성 유지에 효과를 보인다.

[0016] (3) 제안하는 발명은 외부 추가 데이터를 사용하지 않고 훈련 데이터의 형태학적 다양성을 극대화하며, 초거대언어 모델 기반의 데이터 증강 방법에 가깝거나 능가하는 성능을 보인다.

[0017] (4) 제안하는 발명은 적은 양에서도 효과적으로 작동한다.

도면의 간단한 설명

[0018] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 본 발명에서 제안하는 기법(CHEF)에 의한 데이터 증강 결과의 예시를 도시한다.

도 2는 본 발명에서 제안하는 기법(CHEF)의 구조를 도시한다.

도 3은 각 데이터셋의 증강 비율에 따른 제안 기법(CHEF)의 성능 변화를 도시한다.

도 4는 본 발명의 일 실시예에 따른 한국어 데이터 증강 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0019] 본 명세서에 개시되어 있는 본 발명의 개념에 따른 실시예들에 대해서 특정한 구조적 또는 기능적 설명들은 단지 본 발명의 개념에 따른 실시예들을 설명하기 위한 목적으로 예시된 것으로서, 본 발명의 개념에 따른 실시예들은 다양한 형태들로 실시될 수 있으며 본 명세서에 설명된 실시예들에 한정되지 않는다.

[0020] 본 발명의 개념에 따른 실시예들은 다양한 변경들을 가할 수 있고 여러 가지 형태들을 가질 수 있으므로 실시예들을 도면에 예시하고 본 명세서에서 상세하게 설명하고자 한다. 그러나, 이는 본 발명의 개념에 따른 실시예들을 특정한 개시 형태들에 대해 한정하려는 것이 아니며, 본 발명의 사상 및 기술 범위에 포함되는 모든 변경, 균등물, 또는 대체물을 포함한다.

[0021] 제1 또는 제2 등의 용어는 다양한 구성 요소들을 설명하는데 사용될 수 있지만, 상기 구성 요소들은 상기 용어들에 의해 한정되어서는 안 된다. 상기 용어들은 하나의 구성 요소를 다른 구성 요소로부터 구별하는 목적으로만, 예컨대 본 발명의 개념에 따른 권리 범위로부터 벗어나지 않은 채, 제1 구성 요소는 제2 구성 요소로 명명될 수 있고 유사하게 제2 구성 요소는 제1 구성 요소로도 명명될 수 있다.

[0022] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소

에 직접적으로 연결되어 있거나 또는 접속되어 있을 수도 있지만, 중간에 다른 구성 요소가 존재할 수도 있다고 이해되어야 할 것이다. 반면에, 어떤 구성 요소가 다른 구성 요소에 "직접 연결되어" 있다거나 "직접 접속되어" 있다고 언급된 때에는 중간에 다른 구성 요소가 존재하지 않는 것으로 이해되어야 할 것이다. 구성 요소들 간의 관계를 설명하는 다른 표현들, 즉 "~사이에"와 "바로 ~사이에" 또는 "~에 이웃하는"과 "~에 직접 이웃하는" 등도 마찬가지로 해석되어야 한다.

[0023] 본 명세서에서 사용한 용어는 단지 특정한 실시예를 설명하기 위해 사용된 것으로서, 본 발명을 한정하려는 의도가 아니다. 단수의 표현은 문맥상 명백하게 다르게 뜻하지 않는 한, 복수의 표현을 포함한다. 본 명세서에서, "포함하다" 또는 "가지다" 등의 용어는 본 명세서에 기재된 특징, 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것이 존재함을 지정하려는 것이지, 하나 또는 그 이상의 다른 특징들이나 숫자, 단계, 동작, 구성 요소, 부분품 또는 이들을 조합한 것들의 존재 또는 부가 가능성을 미리 배제하지 않는 것으로 이해되어야 한다.

[0024] 다르게 정의되지 않는 한, 기술적이거나 과학적인 용어를 포함해서 여기서 사용되는 모든 용어들은 본 발명이 속하는 기술 분야에서 통상의 지식을 가진 자에 의해 일반적으로 이해되는 것과 동일한 의미를 가진다. 일반적으로 사용되는 사전에 정의되어 있는 것과 같은 용어들은 관련 기술의 문맥상 가지는 의미와 일치하는 의미를 갖는 것으로 해석되어야 하며, 본 명세서에서 명백하게 정의하지 않는 한, 이상적이거나 과도하게 형식적인 의미로 해석되지 않는다.

[0025] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0026] 본 발명에서는, 시퀀스 투 시퀀스 모델(sequence-to-sequence models)의 데이터에서 텍스트를 생성(data-to-text generation)하는 능력에 집중한다. CommonGen(Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. Commongen: A constrained text generation challenge for generative commonsense reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1823-1840.)과 Koran CommonGen(Jaehyung Seo, Seounghoon Lee, Chanjun Park, Yoonna Jang, Hyeonseok Moon, Sugyeong Eo, Seonmin Koo, and Heui-Seok Lim. 2022. A dog is passing over the jet? a text-generation dataset for korean commonsense reasoning and evaluation. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 2233-2249.)에서는 주어진 개념 집합(concept sets)을 기반으로 하는 생성적 상식 추론(generative commonsense reasoning)이 요구된다. 개념 집합은 타당한 문장을 구성할 수 있는 가능성을 결정하는 데 사용된다.

[0027] 본 발명에서는, 이 과정과 완전한 요리(합성 데이터나 문장을 의미함)를 만들기 위해 재료(형태소를 의미할 수 있음)를 결합하는 것 같의 유사성을 느낀다. 특히, Seo 등(2022)은 실질 형태소를 결합하면 한국어 문장 생성을 위한 최적의 구문 분석이 가능하다는 것을 보여준다. 따라서, 본 발명에서는, 형태소 집합의 구성 요소를 조립함으로써, 시퀀스 투 시퀀스 모델의 문장 증강 효과를 활용하는 데 중점을 둔다.

[0028] 제안 기법(CHEF)의 주요 목적은 데이터셋 $\mathcal{D} = \{(\mathrm{inp}^i, l^i)\}_{i=1}^{N}$을 증강하는 것이다. 여기서 $\mathrm{inp}^i = (\mathrm{h}^i, \mathrm{p}^i)$는 텍스트 입력(textual input)을, l^i는 대응하는 레이블(label)을 나타낸다. NLI(Natural Language Inference) 및 STS(Semantic Textual Similarity)와 같이 다중 문장 설정을 고려할 때, h^i를 가설(hypothesis)로 간주하고, 보조 입력 p^i를 전제(premise)로 정의한다. 입력이 단일 문장을 포함한다면(예, YNAT), h^i를 inp^i와 동일한 텍스트로 간주하고, p^i는 빈 문자열(empty string)로 취급할 수 있다. 즉, 본 발명은 단일 문장의 입력이나 다중 문장의 입력 모두에 적용이 가능하다.

[0029] 제안 기법(CHEF)의 구조는 도 2에 도시되어 있다. CHEF는 레이블의 일관성을 유지하도록 돕는 레이블 판별기 모듈(M_D)과 각 형태소 집합을 참조하여 데이터를 증강하는 형태소 조합기(M_G)의 두 요소로 구성된다. 제안 기법에서, 레이블 판별기 모듈(M_D)은 BERT(Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.)와 같은 인코더 아키텍처로 제시되고, 형태소 조합기 모듈은 BART(Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke

Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871-7880, Online. Association for Computational Linguistics.)와 같은 인코더-디코더 아키텍처를 갖는다. 두 모듈 모두, 태스크와 증강될 레이블에 따라, 합성 데이터를 생성하기 위해 학습된다. 이하에서는 각 모듈에 대해 상세히 설명한다.

[0030] **레이블 판별기(Label Discriminator)**

[0031] 레이블 판별기 모듈(M_D)은 데이터 증강 프로세스 동안에 의도된 레이블 일관성을 유지하는 데 도움을 준다. 레이블 판별기 모듈(M_D)은 inp^i를 입력받고 각 레이블로 분류될 확률을 제공하도록 디자인된다. M_D는 소프트맥스(softmax)가 뒤따르는 선형 풀링 레이어(linear pooling layer)를 적용함으로써 inp를 벡터 공간(vector space)으로, 마지막 은닉 상태(hidden state)의 [BOS] 포지션으로, 인코딩한다. 처리된 출력은 $\mathbf{M}_D(\mathbf{inp}^i) \in \mathbb{R}^{1\times n_c}$로 표시한다. 여기서, n_c는 D에서 다루는 태스크를 위한 클래스의 개수를 나타낸다. M_D는 각 inp^i가 l^i로 분류될 확률을 최대화하기 위해 지도 학습된다.

[0032] 간략히 말하면, M_D는, 완전히 문맥화된 문장(full contextualized sentence) inp^i에 의해 획득된 레이블 기대 확률(label expectation probability)을 제공하는, D로 학습된 태스크 모듈이고, 후술할 형태소 조합기의 레이블 교사(label instructor)로 활용된다.

[0033] **형태소 조합기(Morpheme Blender)**

[0034] **성분 준비(Ingredients Preparation)** 형태소 조합기(M_G)의 주된 목적은 주어진 형태소 성분과 목표 레이블을 사용하여 문장을 합성하는 것이다. M_G를 학습하기 위해서, h^i로부터 형태소 성분 $m^i = \{m_1^i, \cdots, m_{n_i}^i\}$을 추출한다. 여기서, n^i는 h^i 내에 포함된 형태소의 개수를 나타낸다. m^i를 준비할 때, 한국어 형태소 분석기(morphological analyzer, 예컨대 mecab-ko)를 이용하여, 한국어는 다양한 기능 형태소를 포함하고 있음을 고려하여, h^i 내의 모든 형태소를 추출한다. 한국어는 동일한 실질 형태소에 대해 정규 및 비정규의 변화를 보인다. 이는, 동일한 실질 형태소 성분을 결합할 수 있는 문장의 다양성이 잠재적으로 높다는 것을 암시한다. 따라서, 실질적인 실질 형태소를 기반으로 형태소 집합을 구축하여, 모델-인지 레시피(model-aware recipe) 내의 형태소를 다양한 어미(endings), 기능 형태소, 및 조사(particles)와 결합할 수 있게 한다.

[0035] 또한, 레이블 l^i를 자연어 형태를 갖는 레이블 워드(label word) v^i로 매핑하는 레이블 매핑 함수 $\psi : l \rightarrow v$를 정의한다. 예컨대, NLI의 예에서, 레이블 "-1"에 대해 "모순(contradiction)", "0"에 대해 "중립(neutral)", "1"에 대해 "수반(entailment)"을 채택할 수 있다.

[0036] 이를 수집하기 위해, 형태소 구성(morpheme ingredients) m^i(p^i가 추가적으로 포함될 수 있음)와 레이블 워드 v^i를 입력으로 사용하는 형태소 조합기를 구축하고, 새롭게 합성된 문장 h'^i을 반환한다. 상세한 학습 과정은 후술한다.

[0037] **티처 포싱(Teacher Forcing)** 구성 합성 능력을 얻기 위하여, M_G는, m^i, p^i, 및 v^i의 연결된 시퀀스(concatenated sequence)를 고려함에 의해, h^i를 생성하도록 학습된다. M_G의 인코더 프레임워크(Enc_G로 표기됨)는 연결된 시퀀스를 수신하고, 입력 내 워드들의 양방향 상호작용(bidirectional interactions)을 포착(capturing)함으로써, 문맥화된 표현(contextualized representation)을 생성한다. 다음으로, M_G는 인코더에 의해 획득된 문맥화된 표현을 활용함으로써 문장을 생성한다. M_G의 시퀀스 투 시퀀스 생성을 학습하기 위한 손실 함수는 수학식 1과 같이 정의된다. […]은 모든 요소의 순차적인 연결(sequentialized concatenation)을 나타내며, 연결된 입력 $seq^i = [m^i, p^i, v^i]$을 정의한다.

[0038] [수학식 1]

[0039] $$\mathcal{L}_{ce} = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{N}\sum_{j}\log P_{\mathrm{M_G}}(h_j^i \mid h_{<j}^i; seq^i)$$

[0040] 시퀀스 투 시퀀스 학습을 통해, $\mathrm{M_G}$는, seq^i 내의 v^i에 의해 부여된 레이블 정보를 약하게 반영하는, 주어진 형태소 집합을 커버하는 전체 문장을 생성할 수 있다.

[0041] **레이블 일관성 지도(Label Consistency Supervision)** 레이블 워드가 형태소 조합기의 입력으로 주어짐에 따라 레이블 정보가 문장 생성 프로세스에서 반영된다 하더라도, $\mathrm{M_G}$의 생성 출력은 여전히 레이블 불일치(label inconsistency) 문제를 갖을 수 있다. 이를 완화하기 위해, 레이블 일관성을 고려하기 위해 사전학습된 판별기 $\mathrm{M_D}$를 활용하는 보조 학습 목적(auxiliary training objective)을 이용한다.

[0042] $\mathrm{M_D}$와 $\mathrm{Enc_G}$의 인코딩된 출력을 정렬함으로써, $\mathrm{Enc_G}$가 해당 형태소 집합 및 태스크별 레이블(task-specific labels)과의 문맥적 관계(contextual relations)를 더 잘 임베딩하고 레이블 일관성을 유지하도록 촉진할 수 있다. 이는 수학식 2와 수학식 3과 같이 정의된 유사도 sim^i를 최대화하는 목적을 학습하는 것으로 이해될 수 있다.

[0043] [수학식 2]

[0044] $$\mathbf{r}_{\mathrm{G}}^i = \mathbf{SoftMax}(\mathbf{Enc_G}(seq^i)\cdot \mathrm{W_G})$$

[0045] [수학식 3]

[0046] $$\mathrm{sim}^i(\mathrm{inp}) = \frac{\mathbf{r}_{\mathrm{G}}^i \cdot \mathbf{M_D}(\mathrm{inp}))}{\|\mathbf{r}_{\mathrm{G}}^i\|\|\mathbf{M_D}(\mathrm{inp}))\|}$$

[0047] $\mathrm{M_D}$의 사전 학습을 통해, $\mathrm{M_D}$의 인코딩된 출력은 $\mathrm{M_G}$가 생성해야 하는 전체 문장 h^i의 레이블 정보를 나타낸다. 보조 학습 목적은 seq^i의 집합을 통해 인코딩된 레이블 표현이 h^i에 의한 레이블 표현과 일치하도록 하는 것이다.

[0048] 직접적인 지도를 위해, 각 $(\mathrm{inp}^i, l^i) \in \mathcal{D}$에 대한 대조 샘플 집합(contrastive sample set) $\mathcal{C}^i \subset \mathcal{D}$, 및 각 레이블 l^i에 대한 대조 레이블 집합 $\mathbf{l}^i = \mathbf{L}\backslash\{l^i\}$를 정의한다. 여기서, $\mathbf{L}$은 $\mathcal{D}$에서 고려되는 모든 가능한 레이블들의 집합을 나타낸다. $\mathcal{C}^i$를 구성하는 과정에서, $\mathbf{l}^i$ 내 각 레이블에 대하여, $\mathcal{D}$로부터 단일 샘플을 랜덤하게 추출한다. 그러면, $\mathrm{M_G}$의 학습을 위한 손실 함수는 수학식 4와 같이 정의된다.

[0049] [수학식 4]

[0050] $$\mathcal{L}_{cont}^i = -\log\frac{\exp(\mathrm{sim}^i(\mathrm{inp}^i)/\tau)}{\sum_{(\mathrm{inp}^j, l^j)\in\mathcal{C}^i}\exp(\mathrm{sim}^i(\mathrm{inp}^j)/\tau)}$$

[0051] 여기서, 수학식 2에 사용된 $\mathrm{W_G}$는 [BOS] 포지션의 인코딩된 표현을 레이블 분류 확률로 매핑하는 선형 풀링 레이어(linear pooling layer)를 나타낸다. 온도 파라미터(temperature parameter) τ는, 더 큰 값은 더 부드러운 분포를, 더 작은 값은 더 뾰족한 분포를 갖도록, 소프트맥스 분포의 샤프니스(sharpness)를 제어한다.

[0052] 이를 요약하면, $\mathrm{M_G}$는 수학식 5 및 수학식 6에 의해 정의되는 $\mathcal{L}_{total}$로 학습된다.

[0053] [수학식 5]

[0054] $$\mathcal{L}_{cont} = -\frac{1}{|\mathcal{D}|}\sum_{i=1}^{N}\mathcal{L}_{cont}^i$$

[0055] [수학식 6]

[0056] $$\mathcal{L}_{total} = (1-\lambda)\mathcal{L}_{ce} + \lambda\mathcal{L}_{cont}$$

[0057] 수학식 5에서, λ는 두 손실 사이의 밸런스 파라미터(balance parameter)로, 가중치를 의미할 수 있다.

[0058] **증강 파이프라인(Augmentation Pipeline)**

[0059] M_G를 활용함에 있어, D 내의 각 (inp^i,l^i)에 대해 단일 데이터를 생성한다. 파이프라인은 다음과 같다.

[0060] 1. inp^i 내의 h^i의 형태소 집합 m^i를 추출.

[0061] 2. $h^i_{aug} = \mathbf{M}_G([m^i, p^i, v^i])$를 생성. 여기서, v^i는 l^i의 레이블 워드를 나타냄.

[0062] 3. h^i_{aug}는, 레이블이 l^i인, D에 대한 증강된 데이터를 나타냄.

[0063] 과도한 증강은 오류 누적과 레이블 혼란을 야기할 수 있기 때문에, 풀샷 러닝(full-shot learning)을 구현함에 있어, D의 작은 부분에 CHEF를 적용할 수 있다.

[0064] 이하에서는, 실험에 이용된 설정을 설명한다.

[0065] **데이터셋(Datasets)**

[0066] 한국어 다중 분류 벤치마크 데이터셋을 활용한다. 각 데이터셋은 학습과 평가에 사용된다.

[0067] **KLUE-NLI** KLUE-NLI 데이터셋(Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021b. Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.)은 명시적으로 자연어 추론(Natural Language Inference, NLI) 태스크를 위해 생성된 데이터셋이다. 학습, 검증, 및 테스트 데이터는 24,998, 3,000, 및 3,000 문장 쌍들을 포함한ㄷ. 이 태스크에서, 모델은 전제(premise)와 가설(hypothesis)이라 불리는 문장 쌍을 처리하고, 수반(entailment), 모순(contradiction), 또는 중립(neutral)인지 추론하여야 한다.

[0068] **KorNLI** KorNLI 데이터셋(Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. Kornli and korsts: New benchmark datasets for korean natural language understanding. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 422-430.)은 또한 한국어 자연어 추론을 위해 디자인되었다. 이 데이터셋은 영어 표준 NLI(SNLI) 및 Multi-Genre NLI (MNLI) 데이터셋과 Cross-lingual NLI (XNLI) 데이터셋을 한국어로 번역함으로써 생성되었다. KorNLI 데이터셋의 학습 데이터는 SNLI와 MNLI 데이터셋으로부터 기계 번역된 942,854 문장 쌍들로 구성되고, 평가 데이터는 XNLI 데이터셋으로부 번역된 7,500 개의 문장 쌍들로 구성된다.

[0069] **KLUE-STS** KLUE-STS 데이터셋(Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, et al. 2021b. Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105.09680.)은 STS(Semantic Textual Similarity) 태스크를 위해 구성되었으며, 학습을 위한 11,668 문장 쌍들, 검증을 위한 519 문장 쌍들, 그리고 테스트를 위한 1,037 문장 쌍들로 구성된다. 이 태스크에서, 모델은 문장 쌍을 평가하고 그들의 의미적 유사성의 정도를 결정한다.

[0070] **KLUE-YNAT** KLUE-YNAT 데이터셋(Park et al., 2021b)는 토픽 분류 테스크를 위해 디자인되었다. 데이터셋은 학습, 검증, 및 테스트 데이터로써, 45,678, 9,107, 및 9,107 샘플들을 포함한다. 이 태스크에서, 모델들은 문장들을 처리하고 미리 정의된 뉴스 카테고리로 할당한다.

[0071] **NSMC** NSMC 데이터셋은 감정 분석 태스크(sentiment analysis task)를 위해 구축되었다. 이는 네이버 플랫폼의 영화 리뷰 및 해당 리뷰의 평점으로부터 도출되었다. NSMC 데이터셋의 학습 데이터는 150,000 리뷰를 포함하고, 테스트 셋은 50,000 리뷰를 포함한다. 이 태스크에서, 모델은 개별 문장을 처리하고 해당 문장의 감정을 판단하여야 한다. 감정은 긍정 또는 부정일 수 있다.

[0072] **모델(Models)**

[0073] 형태소 조합기에서 KoBART를 활용하고, 레이블 판별기에서는 KLUE-BERT-base를 활용한다. KoBART는 작은 모델 파라미터(124M)에서도 탁월한 한국어 텍스트 생성 능력을 보인다(Seo et al., 2022). KLUE-BERT-base를 선택할 수 있는데, 이 모델은 검토된 대안 중에서 가장 소형이며 모델 크기로 인한 판별 프로세스 중 발생할 수 있는

증류 효과를 완화한다(Park et al., 2021b). CHEF의 효과를 검증하기 위해, $BERT^{K}_{Base}$ (KLUE-BERT-base), $RoBERTa^{K}_{Base}$ (KLUE-RoBERTa-base), 및 $RoBERTa^{K}_{Large}$ (KLUE-RoBERTa-large)와 같은 한국어 언어 이해 모델을 선택하였다.

[0074] **비교 방법(Compared Methods)**

[0075] 실험은 풀샷 러닝(full-shot learning)과 퓨샷 러닝(few-shot learning)을 기반으로 수행되었다. BackTranslation (BT) (Sennrich et al., 2016) 및 Korean-EDA (KoEDA) (Wei and Zou, 2019)를 사용한 비교 실험을 진행한다. 또한, GPT-3.5 및 GPT-4를 백본 증강 모델로 활용한다.

[0076] 이하에서는, 실험 결과를 상세히 설명한다.

[0077] **풀샷 학습(Full-shot Learning)**

[0078] 표 1에 나타난 바와 같이, CHEF의 효과와 성능 개선의 정도를, 풀샷 설정에서, 다른 데이터 증강 기법과 비교한다. 학습 데이터셋의 크기에 비례하게 데이터를 증강하였으며, 각 레이블에 대해 학습 데이터의 1%에 해당하는 양이 추가되도록 하였다.

[0079] [표 1]

Model	KLUE-NLI (Acc.)	KorNLI-MNLI (Acc.)	KorNLI-SNLI (Acc.)	KLUE-STS (Pearsonr)	KLUE-YNAT (F1)	NSMC (Acc.)
$BERT^{K}_{Base}$	81.53	80.63	71.56	90.85	85.73	90.3
+KoEDA	81.50	80.12	69.78	90.8	86.01	90.42
+BT	81.30	80.08	70.71	90.65	85.15	90.35
+GPT-3.5+CA	78.90	79.87	68.52	89.49	84.36	89.98
+GPT-3.5+PARA	80.40	80.00	69.42	91.24	86.32	90.45
+GPT-4+CA	80.73	80.60	65.26	90.92	85.50	90.45
+GPT-4+PARA	80.90	80.18	68.59	**91.44**	86.14	90.43
+CHEF	**82.63**	**81.07**	**72.06**	91.26	**86.68**	**90.52**
$RoBERTa^{K}_{Base}$	84.83	80.84	71.84	92.50	85.07	90.08
+KoEDA	85.23	81.27	70.49	92.94	85.18	90.71
+BT	84.53	80.78	69.78	92.68	84.47	90.65
+GPT-3.5+CA	84.90	81.11	69.72	91.16	83.93	90.46
+GPT-3.5+PARA	84.60	80.58	70.12	92.95	85.11	90.87
+GPT-4+CA	84.83	**81.85**	62.13	92.91	84.73	90.83
+GPT-4+PARA	85.03	81.23	70.08	92.99	85.47	**90.94**
+CHEF	**86.0**	81.77	**72.22**	**93.05**	**86.18**	90.74
$RoBERTa^{K}_{Large}$	89.17	81.73	73.26	93.35	85.69	91.28
+KoEDA	89.63	82.24	72.52	93.24	85.32	91.12
+BT	33.33	80.36	72.64	92.89	85.24	91.05
+GPT-3.5+CA	88.03	82.96	33.33	90.73	84.23	33.72
+GPT-3.5+PARA	88.33	83.12	71.86	93.28	85.88	91.31
+GPT-4+CA	89.10	**83.48**	68.94	93.20	85.46	67.97
+GPT-4+PARA	90.13	82.88	72.87	93.51	**86.19**	91.21
+CHEF	**90.47**	82.76	**73.56**	**93.61**	86.12	**91.38**

[0080]

[0081] LLMs을 활용하는 증강 방법은 높은 다양성과 우수한 질적 특성을 가진 문장들을 생성하는 것으로 나타났다. 그러나, 때때로 2%까지의 성능 감소가 나타났다. 이러한 감소는 전반적인 데이터 분포의 상당한 변경으로 인해 발생할 수 있다. 반사실적 증강(counterfactual augmentation)의 경우, chain-of-thought 프롬프트가 제공되더라도, 레이블 변환의 불안정성이 심화된다. 오리지널 문장과 레이블은 형식 준수를 위한 검토 없이 반사실적 변화(counterfactual changes)를 겪어, 변이(variability)가 증가하여 증강 프로세스의 효과를 저해한다.

[0082] BT는 동일한 레이블의 보존을 보장할 수 없어 성능 감소 사례가 발생한다. KoEDA는 데이터를 분할하기 위해 형태소 분석기활용하며 한국어에 대한 동의어 서전(thesaurus)을 활용한다. 이러한 접근법은 적절한 성능 향상을 보이며, 한국어의 특성과 일치하는 형태소 기반 데이터 증강의 효과를 검증한다. 그러나, 레이블 일관성을 보장하는 문제와 문맥적 이해의 부재로 인하여, 성능 향상에 제한이 있다. CHEF가 가장 중요한 성능 향상을 보이는 것을 관찰하였다.

[0083] **증강 사이즈의 변화(Changes in Augmentation Size)**

[0084] 도 3은 각 데이터셋의 증강 비율에 따른 제안 기법(CHEF)의 성능 변화를 도시한다. 합성 데이터의 양이 증가함에 따라,부정적인 노이즈 누적을 초래하는 데이터가 포함될 확률이 증가한다. 풀샷 설정에서, 합성 데이터 샘플의 수를 단순히 확장하는 것이 성능 향상을 보장하지는 않는다. 그러나, CHEF는 데이터 증강의 효과를 유지하며, 데이터 증강이 10% 이내로 제한된 대부분의 경우에서 베이스라인에 비해 우순한 향상을 보인다.

[0085] **퓨샷과 합성 데이터만의 이용(Few-shot and Synthetic-only)**

[0086] NLI 태스크에서 CHEF의 효과를 더 자세히 조사한다. NLI 태스크는 상대적으로 변덕스러운 데이터 증강 효과를 보인다. 32 개의 샘플로 퓨샷 설정에서 비교 실험을 진행하고 각 레이블에 32 개의 증강 문장을 추가하였다. 표 2는 세 가지 다른 시드(seeds)에 대한 결과를 평균화 및 최대값으로 보여준다. LLMs을 사용한 데이터 증강 방법은 풀샷 설정보다 낮은 자원 환경에서 더 효과적이다. CHEF는 제한된 데이터 상황에서도 모델이 성능을 크게 향상시킨다. 또한, CHEF에 의해 생성된 합성 데이터만을 사용하여 학습한 경우의 효과를 평가한다. 결과는 성능의 다양한 향상을 나타내며, CHEF 증강의 품질을 인증한다.

[0087] [표 2]

Model	KLUE-NLI (Acc.)	KorNLI-MNLI (Acc.)	KorNLI-SNLI (Acc.)
$BERT^{K}_{Base}$	37.6/40.1	40.5/41.7	37.5/40.0
+KoEDA	43.0/45.0	43.5/44.9	36.8/38.6
+BT	41.1/43.0	42.9/44.8	38.0/40.2
+GPT-3.5+CA	37.1/37.7	41.6/44.6	38.2/39.5
+GPT-3.5+PARA	41.5/43.4	42.9/44.3	38.7/39.7
+GPT-4+CA	41.7/43.0	46.3/48.5	40.5/41.8
+GPT-4+PARA	44.7/46.1	43.5/45.7	37.7/40.7
+CHEF	**46.5/47.6**	**47.7/49.2**	**41.1/42.3**
+CHEF-SynOnly	41.7/42.5	46.8/47.4	35.1/37.1
$RoBERTa^{K}_{Base}$	37.6/40.1	36.5/36.7	36.1/36.7
+KoEDA	41.2/44.4	38.4/39.9	36.3/36.5
+BT	39.9/42.2	38.9/41.3	36.2/37.1
+GPT-3.5+CA	36.1/36.9	39.1/42.6	36.1/37.1
+GPT-3.5+PARA	39.8/41.8	39.4/40.7	36.1/36.7
+GPT-4+CA	40.3/42.3	41.0/43.3	39.6/41.4
+GPT-4+PARA	**43.3/45.8**	39.4/41.7	39.4/41.6
+CHEF	41.2/43.3	**43.9/45.8**	**40.1/44.7**
+CHEF-SynOnly	38.7/39.5	43.0/45.1	33.9/34.8
$RoBERTa^{K}_{Large}$	47.7/51.7	40.8/44.4	37.5/39.4
+KoEDA	51.2/52.6	43.6/47.0	38.9/40.9
+BT	54.4/56.3	43.5/48.7	38.0/38.8
+GPT-3.5+CA	41.4/43.9	41.7/45.9	38.5/40.6
+GPT-3.5+PARA	52.6/53.7	43.8/47.1	38.5/39.8
+GPT-4+CA	45.6/51.0	46.1/48.8	41.4/ 44.3
+GPT-4+PARA	**56.9/59.5**	45.3/48.2	38.6/41.2
+CHEF	56.5/57.2	**51.5/54.5**	**44.9/47.6**
+CHEF-SynOnly	50.0/57.1	49.8/**54.5**	33.8/34.2

[0088]

[0089] **절제 연구(Ablation Study)**

[0090] 레이블 일관성의 중요성을 명확히 평가하기 위해, CHEF 내에서 판별기 내의 레이블 판별기 요소를 체계적으로 변경하면서 절제 연구를 진행하였다. 판별기에서의 사전학습을 제외하거나 배제함으로써 생기는 성능의 변화는 표 3에 도시되어 있다. 레이블 판변기를 제거하면, 풀샷 설정에서 증강 효과가 감소하는 것을 관찰할 수 있다. 특히, 판별기에 대한 사전학습 단계 없이도 CHEF는 성능을 향상시킬 능력을 유지한다. 형태소 조합기와 사전학습된 레이블 판별기를 대조 손실을 사용하여 통합함으로써, CHEF는 최고 수준의 성능을 달성한다. 이러한 경험적인 결과는 레이블 일관성을 보장하기 위해 판별기를 통합하는 효과를 입증한다.

[0091] [표 3]

Model	KLUE-NLI (Acc.)	KLUE-STS (Pearsonr)	KLUE-YNAT (F1)
$BERT^{K}_{Base}$	81.53	90.85	85.73
$+M_G$	81.83	90.62	86.02
$+M_G+M_D$	82.54	91.05	86.34
$+M_G+M_D$+PT	**82.63**	**91.26**	**86.68**

[0092]

[0093] 도 4는 본 발명의 일 실시예에 따른, 한국어 데이터 증강 방법을 설명하기 위한 흐름도이다.

[0094] 도 4를 참조하면, 한국어 데이터 증강 방법은 적어도 프로세서(Processor) 및/또는 메모리(Memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 다시 말하면, 한국어 데이터 증강 방법을 이루는 단계들 중 적어도 일부는 컴퓨팅 장치에 포함되는 프로세서의 동작으로 이해될 수 있고, 컴퓨팅 장치는 한국어 데이터 증강 장치로 명명될 수 있다. 컴퓨팅 장치는 PC(Personal Computer), 서버(Server), 랩탑 컴퓨터, 태블릿 PC 등을 포함할 수 있다. 또한, 컴퓨팅 장치는 물리적으로 분리된 복수의 장치들에 의해 구현되는 것도 가능하며, 실시예에 따라 클라우드 환경에서 구현되는 것도 가능하다. 이하에서, 한국어 데이터 증강 방법을 설명함에 있어, 앞선 기재와 중복되는 내용에 관하여는, 그 구체적인 기재를 생략하기로 한다.

[0095] 우선, 레이블 판별 모델이 생성된다(S110). 레이블 판별 모델은 소정의 학습 데이터를 이용하여, 문장(입력)에 해당하는 레이블 확률을 추론하도록 소정의 신경망 모델을 학습함으로써 생성될 수 있다. 따라서, 학습 데이터는 문장과 해당 문장의 레이블을 포함할 수 있다. 학습 모델은 디코더 구조의 신경망 모델로써, 예시적인 학습 모델은 BERT 기반의 모델일 수 있다.

[0096] 형태소 조합 모델이 생성된다(S120). 형태소 조합 모델은 소정의 학습 데이터를 이용하여, 형태소 집합과 레이블 워드를 입력으로 받아 새로운 문장을 생성하도록 소정의 신경망 모델을 학습함으로써 생성될 수 있다. 따라서, 학습 데이터는 문장에 포함된 형태소 집합(실질 형태소의 집합을 의미할 수 있음)과 문장의 레이블에 대응하는 레이블 워드를 포함할 수 있다. 이때, 형태소 집합은 문장에 대한 형태소 분석을 통하여 생성될 수 있다. 학습 모델은 인코더-디코더 구조의 신경망 모델로써, 예시적인 학습 모델은 BART 기반의 모델일 수 있다.

[0097] 입력 문장(또는 타겟 문장)에 대한 형태소 분석이 수행된다(S130). 이를 통해 입력 문장에 포함되는 형태소를 추출할 수 있다. 형태소 분석은 임의의 형태소 분석 기법을 이용하여 수행될 수 있다.

[0098] 형태소 조합 모델을 이용하여 입력 문장에 대응하는 증강 데이터를 생성할 수 있다(S140). 즉, 형태소 집합과 레이블 워드를 형태소 조합 모델에 입력함으로써, 입력 문장과 동일한 레이블을 갖는 합성 문장이 생성될 수 있다.

[0099] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시예들에서 설명된 장치 및 구성 요소는, 예를 들어, 프로세서, 콘트롤러, ALU(Arithmetic Logic Unit), 디지털 신호 프로세서(Digital Signal Processor), 마이크로컴퓨터, FPA(Field Programmable array), PLU(Programmable Logic Unit), 마이크로프로세서, 또는 명령(instruction)을 실행하고 응답할 수 있는 다른 어떠한 장치와 같이, 하나 이상의 범용 컴퓨터 또는 특수 목적 컴퓨터를 이용하여 구현될 수 있다. 처리 장치는 운영 체제(Operation System, OS) 및 상기 운영 체제 상에서 수행되는 하나 이상의 소프트웨어 애플리케이션을 수행할 수 있다. 또한, 처리 장치는 소프트웨어의 실행에 응답하여, 데이터를 접근, 저장, 조작, 처리 및 생성할 수도 있다. 이해의 편의를 위하여, 처리 장치는 하나가 사용되는 것으로 설명된 경우도 있지만, 해당 기술 분야에서 통상의 지식을 가진 자는, 처리 장치가 복수 개의 처리 요소(Processing Element) 및/또는 복수 유형의 처리 요소를 포함할 수 있음을 알 수 있다. 예를 들어, 처리 장치는 복수 개의 프로세서 또는 하나의 프로세서 및 하나의 콘트롤러를 포함할 수 있다. 또한, 병렬 프로세서(Parallel Processor)와 같은, 다른 처리 구성(Processing Configuration)도 가능하다.

[0100] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 매체 또는 장치, 또는 전송되는 신호 파(Signal Wave)에 영구적으로, 또는 일시적으로 구체화(Embody)될 수 있다. 소프트웨어는 네트워크로 연결된 컴퓨터 시스템 상에 분산되어서, 분산된 방법으로 저장되거나 실행될 수도 있다. 소프트웨어 및 데이터는 하나 이상의 컴퓨터 판독 가능 기록 매

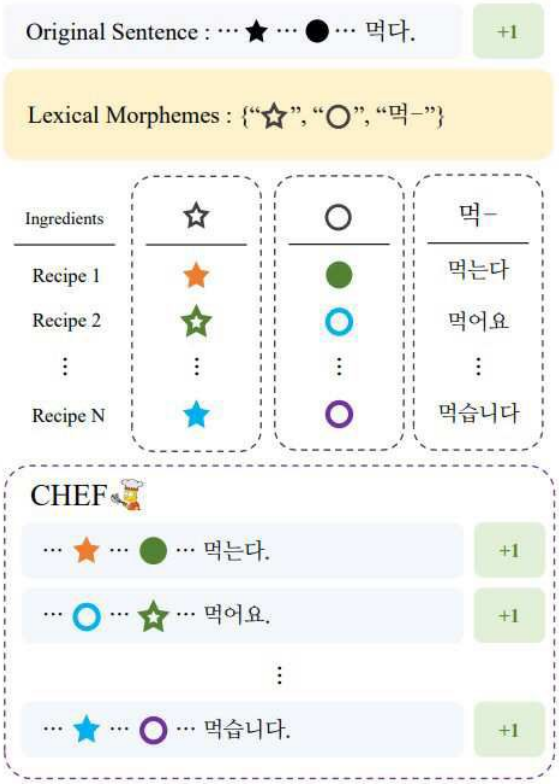
체에 저장될 수 있다.

[0101] 실시예에 따른 방법은 다양한 컴퓨터 수단을 통하여 수행될 수 있는 프로그램 명령 형태로 구현되어 컴퓨터 판독 가능 매체에 기록될 수 있다. 상기 컴퓨터 판독 가능 매체는 프로그램 명령, 데이터 파일, 데이터 구조 등을 단독으로 또는 조합하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시예를 위하여 특별히 설계되고 구성된 것들이거나 컴퓨터 소프트웨어 당업자에게 공지되어 사용 가능한 것일 수도 있다. 컴퓨터 판독 가능 기록 매체의 예에는 하드 디스크, 플로피 디스크 및 자기 테이프와 같은 자기 매체(Magnetic Media), CD-ROM, DVD와 같은 광기록 매체(Optical Media), 플롭티컬 디스크(Floptical Disk)와 같은 자기-광 매체(Magneto-optical Media), 롬(ROM), 램(RAM), 플래시 메모리 등과 같은 프로그램 명령을 저장하고 수행하도록 특별히 구성된 하드웨어 장치가 포함된다. 프로그램 명령의 예에는 컴파일러에 의해 만들어지는 것과 같은 기계어 코드뿐만 아니라 인터프리터 등을 사용해서 컴퓨터에 의해서 실행될 수 있는 고급 언어 코드를 포함한다. 상기된 하드웨어 장치는 실시예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

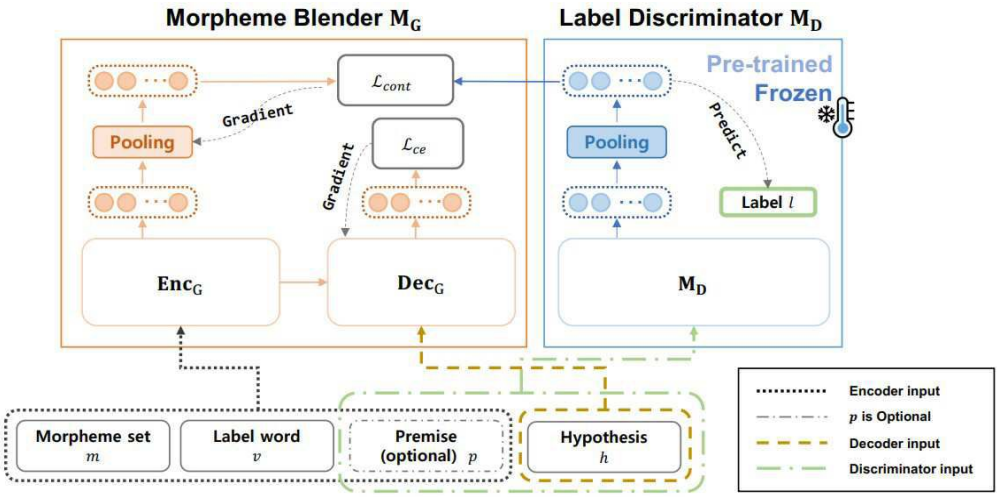
[0102] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

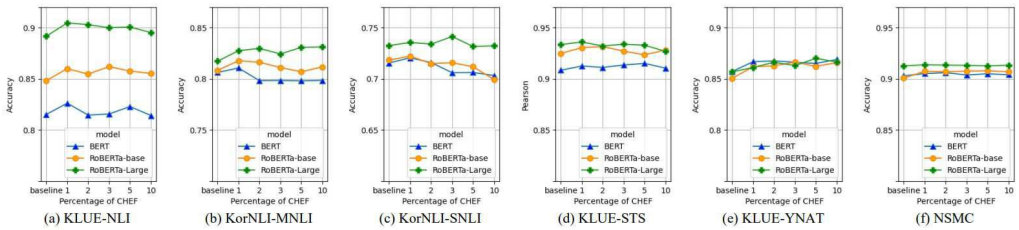
도면1



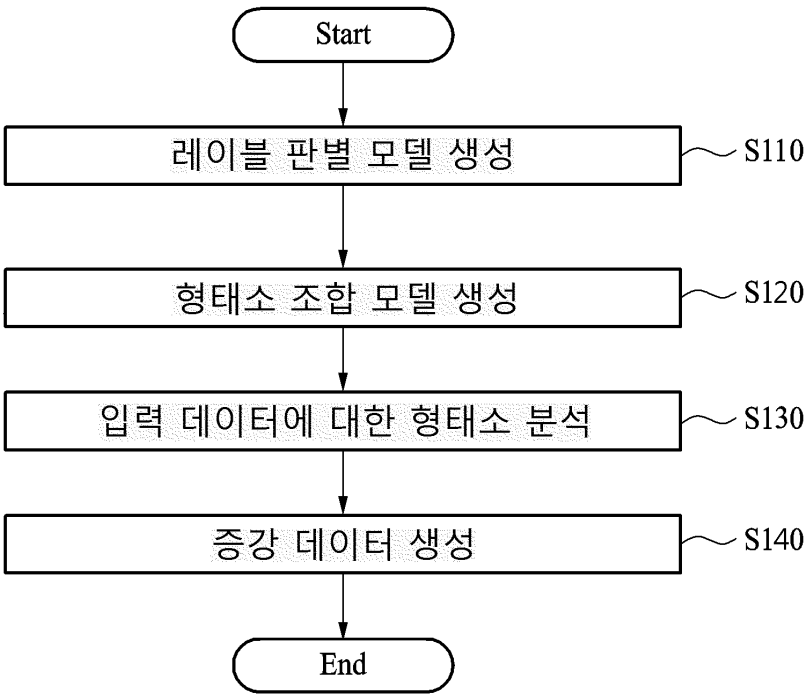
도면2


Morpheme Blender M_G
Label Discriminator M_D
$\mathcal{L}_{cont}$
$\mathcal{L}_{ce}$
Gradient
Gradient
Pooling
Pooling
Pre-trained
Frozen
Predict
Label l
Enc_G
Dec_G
M_D
Morpheme set m
Label word v
Premise (optional) p
Hypothesis h
Encoder input
p is Optional
Decoder input
Discriminator input

도면3


(a) KLUE-NLI
(b) KorNLI-MNLI
(c) KorNLI-SNLI
(d) KLUE-STS
(e) KLUE-YNAT
(f) NSMC

도면4


Start
레이블 판별 모델 생성
S110
형태소 조합 모델 생성
S120
입력 데이터에 대한 형태소 분석
S130
증강 데이터 생성
S140
End