(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(45) 공고일자 2023년10월24일
(11) 등록번호 10-2593448
(24) 등록일자 2023년10월19일

(51) 국제특허분류(Int. Cl.)
G06F 40/58 (2020.01) *G06F 40/205* (2020.01)
G06F 40/237 (2020.01) *G06F 40/279* (2020.01)
G06N 3/08 (2023.01)
(52) CPC특허분류
G06F 40/58 (2020.01)
G06F 40/205 (2020.01)
(21) 출원번호 10-2022-0151593
(22) 출원일자 2022년11월14일
심사청구일자 2022년11월14일
(56) 선행기술조사문헌
Banon, Marta, et al., ParaCrawl: Web-scale acquisition of parallel corpora., Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics., 2020*
Junczys-Dowmunt, Marcin., Dual conditional cross-entropy filtering of noisy parallel corpora., arXiv preprint arXiv:1809.00197, 2018*
Axelrod, Amittai, Anish Kumar, and Steve Sloto., Dual monolingual cross-entropy delta filtering of noisy parallel data., Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared T*
Ippolito, Daphne., Understanding the Limitations of Using Large Language Models for Text Generation., Doctoral dissertation, Google, 2021*
*는 심사관에 의하여 인용된 문헌

(73) 특허권자
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스 애기능생활관 311호 NLP&AI연구실
문현석
서울특별시 강남구 자곡로 260, 422동 903호
(74) 대리인
김등용

전체 청구항 수 : 총 3 항 심사관 : 김영신

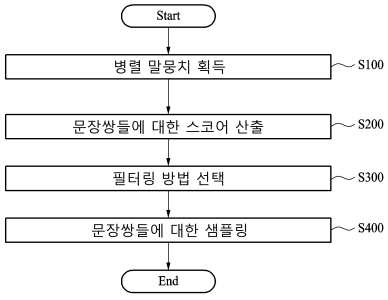
(54) 발명의 명칭 **의미적 유사도 기반 병렬 말뭉치 정제 방법 및 장치**

(57) 요 약

의미적 유사도에 기반한 병렬 말뭉치 정제 방법 및 장치가 개시된다. 상기 병렬 말뭉치 정제 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 병렬 말뭉치 정제 방법으로서, 각각이 소스 문장과 타겟 문장을 포함하는 문장쌍들을 포함하는 병렬 말뭉치를 획득하는 단계, 복수의 필터링 기법들 각각에 대하여, 상기 문장쌍

(뒷면에 계속)

대 표 도 - 도2



들 각각에 대한 스코어를 산출하는 단계, 상기 복수의 필터링 기법들 중에서 어느 하나의 필터링 기법을 선택하는 단계, 및 상기 어느 하나의 필털이 기법에 의한 스코어를 기초로 상기 문장쌍들을 샘플링하는 단계를 포함한다.

(52) CPC특허분류
G06F 40/237 (2022.01)
G06F 40/279 (2020.01)
G06N 3/08 (2023.01)

이 발명을 지원한 국가연구개발사업
과제고유번호 1345349433
과제번호 2021R1A6A1A03045425
부처명 교육부
과제관리(전문)기관명 한국연구재단
연구사업명 이공학학술연구기반구축
연구과제명 Human-inspired AI 연구소
기 여 율 1/2
과제수행기관명 고려대학교
연구기간 2022.03.01 ~ 2023.02.28
이 발명을 지원한 국가연구개발사업
과제고유번호 1711159611
과제번호 2020-0-01819-003
부처명 과학기술정보통신부
과제관리(전문)기관명 정보통신기획평가원
연구사업명 정보통신방송혁신인재양성
연구과제명 ICT명품인재양성(고려대학교)
기 여 율 1/2
과제수행기관명 고려대학교산학협력단
연구기간 2022.01.01 ~ 2022.12.31
공지예외적용 : 있음

명 세 서

청구범위

청구항 1

적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 병렬 말뭉치 정제 방법에 있어서,

각각이 소스 문장과 타겟 문장을 포함하는 문장쌍들을 포함하는 병렬 말뭉치를 획득하는 단계;

복수의 필터링 기법들 각각에 대하여, 상기 문장쌍들 각각에 대한 스코어를 산출하는 단계;

상기 복수의 필터링 기법들 중에서 어느 하나의 필터링 기법을 선택하는 단계; 및

상기 어느 하나의 필터링 기법에 의한 스코어를 기초로 상기 문장쌍들을 샘플링하는 단계를 포함하고,

상기 복수의 필터링 기법들은 제1 필터링 기법을 포함하고, 상기 제1 필터링 기법은 NMT(Neural Machine Translation, 신경망 기계 번역) 모델을 이용하여 상기 문장쌍들 각각에 대한 스코어를 산출하고,

상기 복수의 필터링 기법들은 제2 필터링 기법을 포함하고, 상기 제2 필터링 기법은 PLM(Pre-trained Language Model)을 이용하여 상기 문장쌍들 각각에 대한 스코어를 산출하고,

상기 복수의 필터링 기법들은 제3 필터링 기법을 포함하고, 상기 제3 필터링 기법은 문장 임베딩(sentence embedding)을 이용하여 상기 문장쌍들 각각에 대한 스코어를 산출하고,

상기 어느 하나의 필터링 기법을 선택하는 단계는,

학습 대상인 기계 번역이 미리 정해져 있는 경우 상기 제1 필터링 기법을 선택하고,

산출된 스코어의 평균이 미리 정해진 제1 임계값 보다 작은 경우 제2 필터링 기법을 선택하고,

산출된 스코어의 평균이 미리 정해진 제2 임계값 보다 큰 경우 제3 필터링 기법을 선택하는,

병렬 말뭉치 정제 방법.

청구항 2

삭제

청구항 3

삭제

청구항 4

삭제

청구항 5

삭제

청구항 6

제1항에 있어서,

상기 샘플링 하는 단계는,

선택된 필터링 기법에 의한 스코어 값을 기초로, 가장 높은 스코어를 갖는 a%(a는 임의의 실수)의 문장쌍들을 제외하고, 다음으로 높은 스코어를 갖는 b%(b는 임의의 실수)의 문장쌍들을 추출하여 학습 데이터를 구축하는,

병렬 말뭉치 정제 방법.

청구항 7

제1항에 있어서,

상기 샘플링 하는 단계는,

선택된 필터링 기법에 의한 스코어 값을 기초로, 중간값을 갖는 문장쌍을 기준으로 상하로 c%(c는 임의의 실수)의 문장쌍들을 추출하여 학습 데이터를 구축하는,

병렬 말뭉치 정제 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 인공신경망 기반 번역 시스템에 관한 것으로, 특히 번역 시스템 생성을 위해 요구되는 학습 데이터 구축 방법 및 장치에 관한 것이다.

배 경 기 술

[0002] 기계번역(Machine Translation, MT)은 전문가가 세워 놓은 번역 규칙 및 통계적 모델링 기법을 거쳐, 최근에는 인공 신경망 기술로 발전해왔다. 특히, 인공 신경망 기반 번역 기술은 이전 기술과는 비교할 수 없을 정도로 높은 번역 정확도를 보이며 현재 대부분의 분야에서 채용되고 있다.

[0003] 하지만, 인공 신경망 기반 번역 모델의 성능은 훈련 말뭉치(training corpus)의 품질에 매우 민감하게 반응한다. 이는 아무리 많은 번역 말뭉치를 활용하더라도 말뭉치 내에 섞인 저품질의 데이터들에 의해, 훈련된 번역 모델의 성능이 크게 감소하는 것을 의미한다. 이러한 현상은 통계기반 번역 모델링을 활용할 때보다, 인공신경망 기반 번역 모델을 활용할 때 더 명확하게 드러난다.

[0004] 실제로 말뭉치의 품질을 향상시키려는 연구는 많은 주목을 받으며 국내외에서 활발하게 이루어지고 있다. 특히, 말뭉치 정제(filtering)에 요구되는 번역 전문 인력의 개입을 최소화하고 이를 인공지능 기술로 대체하려는 시도는 현재 많이 이루어지고 있는 실정이다.

선행기술문헌

특허문헌

[0005] (특허문헌 0001) 대한민국 공개특허 제2022-0033652호 (2022.03.17. 공개)

(특허문헌 0002) 대한민국 등록특허 제2395811호 (2022.05.09. 공고)

(특허문헌 0003) 대한민국 공개특허 제2019-0060227호 (2019.06.03. 공개)

(특허문헌 0004) 대한민국 등록특허 제2122081호 (2020.06.26. 공고)

발명의 내용

해결하려는 과제

[0006] 현재, 다양한 말뭉치 정제 기술이 등장하였고, 이들 기술은 좋은 성능을 내고 있다. 그러나, 일관적으로 좋은 성능을 내는 기술은 아직 정립되지 않았다. 특히, 이러한 기술들은 제한된 언어쌍에 대해서만 발명되고 활용되고 있을 뿐, 한국어와 관련된 언어쌍에서 이러한 기술이 적용된 사례는 아직까지 많이 존재하지 않는다. 특정 말뭉치에서 좋은 성능을 내는 말뭉치 정제기술이라고 해서, 해당 기술이 모든 말뭉치에 범용적으로 적용 가능한 기술은 아니다. 이는 인공 신경망 기반 기술이 대부분 도메인에 매우 큰 영향을 받기 때문이다. 또한, 이는 고정된 정제 기술 한가지를 활용하는 것이 아니라, 활용하려는 말뭉치와 정제하기 위한 목적에 따라 다른 정제 기술을 적용하는 것이 올바른 방향이 될 수 있음을 보여준다.

[0007] 이에, 본 발명에서는, 한국어를 포함하는 언어쌍에서 좋은 성능을 발휘하는 병렬 말뭉치 정제 장치 및 방법을 제안하고자 한다.

과제의 해결 수단

[0008] 본 발명의 일 실시예에 따른 병렬 말뭉치 정제 방법은 적어도 프로세서를 포함하는 컴퓨팅 장치에 의해 수행되는 병렬 말뭉치 정제 방법으로서, 각각이 소스 문장과 타겟 문장을 포함하는 문장쌍들을 포함하는 병렬 말뭉치를 획득하는 단계, 복수의 필터링 기법들 각각에 대하여, 상기 문장쌍들 각각에 대한 스코어를 산출하는 단계, 상기 복수의 필터링 기법들 중에서 어느 하나의 필터링 기법을 선택하는 단계, 및 상기 어느 하나의 필털이 기법에 의한 스코어를 기초로 상기 문장쌍들을 샘플링하는 단계를 포함한다.

발명의 효과

[0009] 본 발명의 실시예에 따른 병렬 말뭉치 정제 방법 및 장치에 의할 경우, 한국어와 관련된 언어쌍에 대한 고품질의 병렬 말뭉치를 생성할 수 있는 효과가 있다.

도면의 간단한 설명

[0010] 본 발명의 상세한 설명에서 인용되는 도면을 보다 충분히 이해하기 위하여 각 도면의 상세한 설명이 제공된다.

도 1은 필터링 방법, 즉 스코어링 함수에 따른 성능 평가 표가 도시되어 있다.

도 2는 본 발명의 일 실시예에 따른 병렬 말뭉치 정체 방법을 설명하기 위한 흐름도이다.

발명을 실시하기 위한 구체적인 내용

[0011] 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.

[0012] 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.

[0013] 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.

[0014] 어떤 구성 요소가 다른 구성 요소에 "연결되어" 있다거나 "접속되어" 있다고 언급된 때에는, 그 다른 구성 요소에 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.

[0015] 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.

[0016] 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.

[0017] 이하, 본 명세서에 첨부된 도면들을 참조하여 본 발명의 실시예들을 상세히 설명한다. 그러나, 특허출원의 범위

가 이러한 실시예들에 의해 제한되거나 한정되는 것은 아니다. 각 도면에 제시된 동일한 참조 부호는 동일한 부재를 나타낸다.

[0018] 기계 번역의 학습을 위해서는 병렬 말뭉치(parallel corpus)가 필요하다. 딥러닝 프레임워크를 고려할 때, 흔히 데이터셋의 양에 따라 고성능의 번역 모델이 생성될 것으로 기대된다. 그러나, 최근 연구에서는, 데이터의 양보다는 질이 더욱 중요하다는 것이 밝혀졌다. 이에, 본 발명에서는, 의미적 유사도 기반의 병렬 말뭉치 정제 방법 및 장치를 제안한다.

[0019] 본 발명에서는 1) "병렬 말뭉치 정제(Parallel Corpus Filtering, PCF)를 효율적으로 만드는 것은 무엇인가?"와 2) "스코어 함수(scoring function)는 신뢰할 수 있는가?"의 두 가지 물음에 대한 검증에 집중한다. 이를 위해, WMT 접근법을 포함한 다양한 딥러닝 기반 PCF 기법들을 채용한다. WMT 세팅을 채용하여, 정제된 데이터셋으로 NMT 모델을 학습하고 NTM 모델의 성능을 고려할 수 있다. 다양한 스코어 함수를 적용하여, 상이한 환경에서 스코어 함수를 효율적으로 만드는 요인들을 파악할 수 있다.

[0020] 스코어링 방법(Scoring Method)

[0021] 1. NMT 모델 기반 방법

[0022] 대표적인 스코어링 기법은 NMT(Neural Machine Translation) 모델을 이용하는 것이다. NMT 모델 기반 방법은 제1 스코어링 방법 또는 제1 필터링 방법 등으로 명명될 수 있다. NMT 모델 기반 방법은 쌍을 이루는 두 문장들 사이의 의미적 일관성(semantical coherence)을 평가하는 스코어링 방법으로 간주될 수 있다. 특히, 소스 문장(src)과 타겟 문장(tgt)을 포함하는 문장 쌍(src, tgt)과 정제된 코퍼스(clean corpus)로 학습된 NMT 모델들(θ_{s2t} 및 θ_{t2s})에 대하여, 각 번역 모델에 대한 크로스 엔트로피 스코어(cross entropy score)가 산출될 수 있다. 크로스 엔트로피 스코어는 수학식 1과 같다. 또한, NMT 모델 θ_{s2t}는 소스 문장을 타겟 문장으로 번역하는 번역 모델을, NMT 모델 θ_{t2s}는 타겟 문장을 소스 문장으로 번역하는 번역 모델을 의미할 수 있다.

[0023] [수학식 1]

$$\mathrm{TR}_{s2t} = \frac{1}{|tgt|}\sum_{i}\log P(tgt_i|src, tgt_{<i}, \theta_{s2t})$$

$$\mathrm{TR}_{t2s} = \frac{1}{|src|}\sum_{i}\log P(src_i|tgt, src_{<i}, \theta_{t2s})$$

[0024]

[0025] 수학식 1에서, src_i와 tgt_i는 각각 소스 문장(src)과 타겟 문장(tgt) 내의 각 토큰을 나타낸다. 그런 다음, 두 개의 크로스 엔트로피 스코어를 더함으로써, 번역 품질을 평가하는 인덱스로써의 스코어 $score_{\mathrm{TRs}}$가 산출될 수 있다.

[0026] 실시예에 따라, Junczys-Dowmunt의 제안(Marcin Junczys-Dowmunt. 2018. Dual conditional cross-entropy filtering of noisy parallel corpora. arXiv preprint arXiv:1809.00197.)과 같이, 스코어 $score_{\mathrm{TRs_dual}}$를 채용할 수 있다. 이는 두 문장들 사이의 질적 차이(quality difference)를 반영하고, 이의 목적은 큰 질적 차이를 갖는 문장쌍의 퀄리티 스코어(quality score)를 낮추기 위함이다. 이러한 두 가지의 스코어는 수학식 2와 같이 정의될 수 있다.

[0027] [수학식 2]

$$score_{\mathrm{TRs}} = \mathrm{TR}_{s2t} + \mathrm{TR}_{t2s}$$

$$score_{\mathrm{TRs\_dual}} = \frac{1}{2}(\mathrm{TR}_{s2t} + \mathrm{TR}_{t2s}) + |\mathrm{TR}_{s2t} - \mathrm{TR}_{t2s}|$$

[0028]

[0029] 본 발명에서는, 문장쌍(sentence pairs)의 의미적 유사도(semantic similarity)에 기반한 스코어링 방법으로써, NMT 모델을 이용한 스코어링 방법을 정의한다. 그러나, 번역 모델의 성능과 경향은 학습 말뭉치의 특성에 따라 변화하기 때문에, 학습 코퍼스에 편향될 수 있는 문제점이 존재한다.

[0030] 2. PLM(Pre-trained Language Model) 기반 방법

[0031] 말뭉치쌍의 품질은 문장쌍의 의미적 유사도를 통해서 뿐만아니라, 문장 자체의 유창성(fluency)를 통해서도 평

가될 수 있다. PLM 기반 방법은 제2 필터링 방법 또는 제2 스코어링 방법 등으로 명명될 수 있다. 문장들의 유창성은 PLM(Pre-trained Language Model)을 통해 체크될 수 있다(Amittai Axelrod, Anish Kumar, and Steve Sloto. 2019. Dual monolingual cross-entropy delta filtering of noisy parallel data. In Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2), pages 245-251.). 특히, GPT(또는 GPT 2)를 이용함으로써 수행될 수 있다(Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.). GPT2는 입력으로 이전 토큰들을 수신하고 다음 토큰을 생성하도록 학습된다. 이러한 학습의 목적은 문장의 타당성(plausibility)과 함께 의미적 유창성(semantical fluency)에 대한 평가를 가능케 한다.

[0032] 본 발명에서는, 사전학습된 언어 모델(예컨대, GPT2)을 이용하여 주어진 문장쌍(src, tgt)에 대한 크로스 엔트로피를 도출하는 기법을 제안한다. 이를 통해 문장쌍의 품질을 평가할 수 있다. 제안하는 크로스 엔트로피는 수학식 3과 같다.

[0033] [수학식 3]

$$\mathrm{LM}_{src} = \frac{1}{|src|}\sum_{i}\log P(src_i|src_{<i},\theta_{src})$$

$$\mathrm{LM}_{tgt} = \frac{1}{|tgt|}\sum_{i}\log P(tgt_i|tgt_{<i},\theta_{tgt})$$

[0034]

[0035] 수학식 3에서, src_i와 tgt_i는 각각 src와 tgt 내의 각 토큰을 나타낸다. 두 문장들의 크로스 엔트로피를 더함으로써, LMs 스코어가 도출될 수 있다. src 문장과 tgt 문장이 모두 잘 생성되었다면, 대응되는 문장쌍은 고품질로 평가될 수 있다.

[0036] 실시예에 따라, 두 스코어를 단순히 더하는 것 대신 스코어 상에서 두 문장들의 질적 차이를 반영하는 LMs_dual이라 명명되는 스코어를 채용할 수 있다. 그러나, 본 발명이 이에 제한되는 것은 아니며, 실시예에 따라 다양하게 변형될 수 있다. 상술한 스코어는 수학식 4와 같다.

[0037] [수학식 4]

$$score_{\mathrm{LMs}} = \mathrm{LM}_{src} + \mathrm{LM}_{tgt}$$

$$score_{\mathrm{LMs\_dual}} = \frac{1}{2}(\mathrm{LM}_{src} + \mathrm{LM}_{tgt}) + |\mathrm{LM}_{src} - \mathrm{LM}_{tgt}|$$

[0038]

[0039] 다양한 특징을 갖는 각 말뭉치에 상술한 방법을 적용함으로써, 언어 모델을 통해 측정된 유창성이 필터링에 어떤게 영향을 주는지 검증할 수 있다.

[0040] 3. 문장 임베딩(Sentence Embedding) 기반 방법

[0041] 최근, PCF 연구 분야에서 (다언어(multilingual)) 문장 임베딩(sentence embedding)이 널리 채용되고 있다. 문장 임베딩 기반 방법은 제3 필터링 방법 또는 제3 스코어링 방법 등으로 명명될 수 있다. 문장 임베딩을 이용하는 목적은 공유된 잠재 공간(shard latent space) 내에서 문장의 전체적인 의미를 나타내고, 쌍을 이루는 문장들 사이의 임베딩 유사도(embedding similarity)를 문장쌍의 품질로 간주하려는 것이다. 특히, LASER(Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero shot cross-lingual transfer and beyond. Transactions of the Association for Computational Linguistics, 7:597-610.)와 같은 문장 임베딩이 이용될 수 있다. 상기 (다언어) 임베딩을 사용하는 경우, 주어진 문장쌍(src, tgt)에 대하여, 문장쌍에 대한 질적 인덱스(quality index) $score_{se}$는 수학식 5와 같이 측정될 수 있다.

[0042] [수학식 5]

$$score_{\mathrm{se}} = \frac{e(src)\cdot e(tgt)}{\|e(src)\|\cdot\|e(tgt)\|}$$

[0043]

[0044] 수학식 5에서, $e(\cdot)$는 주어진 문장에 대한 다언어 문장 임베딩을 나타낸다. 본 발명에서는, 널리 채용되는 두 개의 다언어 문장 임베딩으로, LASER와 LABSE(Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. arXiv preprint arXiv:2007.01852.)를 이

용할 수 있다. 이러한 스코어링 방법을 각각 $score_{\text{LASER}}$ 와 $score_{\text{LABSE}}$ 로 표기한다.

[0045] 샘플링 기법(Sampling Methodology)

[0046] "스코어링 함수는 믿을 수 있는가?"라는 질문을 검증하기 위한 샘플링 기법들을 제안한다. 높은 스코어의 문장쌍들을 높은 품질의 데이터로 간주하는 것은 타당하지 않을 수 있다.

[0047] 딥러닝 모델의 성능이 학습 말뭉치의 특성에 따라 결정되기 때문에, 딥러닝 모델의 성능은 데이터의 품질에 대응하지 않을 수 있다. 본 발명에서는 이와 같은 스코어 기반의 데이터 필터링이 모델이 특정 경향에 치우치게 많들 수 있다는 가정 하에, 이를 다양한 샘플링 기법에 대한 실험을 통해 검증하고자 한다. 이는 높은 스코어의 문장들을 높은 품질로 평가하는 방법이 합리적인가를 체크하는 것이다.

[0048] 전체 데이터셋 D가 스코어링 함수에 의해 내림차순으로 정렬되었다고 가정한다. 그리고, t번째 데이터에 대한 특정 스코어 값을 $score_i$ 로 나타낸다. 또한, 각 스코어링 방법을 적용하여 필터링된 데이터셋을 수학식 6과 같이 $D_{score,k}$ 로 나타낸다.

[0049] [수학식 6]

[0050] $$D_{(score,k)} = \text{top k}(D|score)$$

[0051] 수학식 6에서, D는 전체 데이터셋을 score는 이전에 정의된 스코어링 함수를 나타낸다. 수학식 6은, $D_{score,k}$은 높은 스코어 값을 갖는 k개의 문장들을 포함함을 의미한다.

[0052] 1. 탑 샘플링(Top Sampling)

[0053] 탑 샘플링 방법(제1 샘플링 방법으로 명명될 수도 있음)은 종래의 접근법과 동일하게 높은 스코어의 문장을 추출하는 것이다. 탑 샘플링 방법에 의해 구축된 필터링된 데이터셋 D_f 는 수학식 7과 같이 정의될 수 있다.

[0054] [수학식 7]

[0055] $$D_f = D_{(score,k)}$$

[0056] WMT를 포함하는 기존의 PCF 연구들은 이러한 샘플링 기법을 채용한다. 그러나, 본 발명에서는 높은 스코어의 데이터를 높은 품질의 데이터로 고려할 수 있는지에 대한 의문을 제기한다.

[0057] 2. 미들 샘플링(Middle Sampling)

[0058] 미들 샘플링 방법(제2 샘플링 방법으로 명명될 수도 있음)은 필터링된 데이터셋을 구축하기 위하여, 가장 높은 스코어의 데이터를 제외하고 높은 스코어의 문장들을 추출한다. 이러한 샘플링 방법을 적용함에 있어, 본 발명에서는 스코어링 함수를 통해 획득된 가장 높은 스코어는 스코어링 모델 자체에 매우 편중되어 있음을 가정한다. 이러한 샘플링 방법은 수학식 8과 같이 공식화될 수 있다.

[0059] [수학식 8]

[0060] $$n_{top} = \|D\| * 0.05$$
$$D_f = \text{top k}(D \backslash D_{(score,n_{top})}|score)$$

[0061] 이러한 샘플링 방법에 의할 경우, 가장 높은 점수를 갖는 a%(a는 임의의 실수로써, 예시적인 값은 5일 수 있음)의 문장들을 제외하고, 다음으로 높은 스코어를 갖는 b%(b는 임의의 실수)의 문장들이 선택될 수 있다.

[0062] 3. 메디안 샘플링(Median Sampling)

[0063] 메디안 샘플링 방법(제3 샘플링 방법으로 명명될 수도 있음)은 중간값의 스코어 값을 갖는 문장들을 선택하는 것이다. 메디안 샘플링 방법에서는, 높은 스코어의 데이터 보다 중간값의 스코어 값을 갖는 데이터가 추출될 수 있다. 예시적으로, 중간 순위의 스코어로부터 상하위 c%(c는 임의의 실수로써, 예시적인 값은 2.5임)이 데이터가 추출될 수 있다. 메디안 샘플링 방법을 통해 추출된 필터링된 데이터셋은 수학식 9와 같이 정의될 수 있다.

[0064] [수학식 9]

$$score_i^{med} = |i - \frac{\|D\|}{2}|$$

[0065] $$D_f = D_{(score^{med},k)}$$

[0066] 메디안 샘플링 방법은, 높은 스코어의 데이터나 낮은 스코어의 데이터 모드 높은 품질을 보장하지 않음을 암시한다. 이와는 달리, 양 방향으로의 편향을 제거함으로써, 중간값의 스코어를 갖는 데이터가 높은 품질을 나타낼 수 있다. 이러한 샘플링 방법을 통해 필터링된 데이터에 대하여 NMT 모델을 학습함으로써, 데이터의 품질을 평가하는 과정에서 스코어링 함수가 편향되었는지를 체크할 수 있다.

[0067] 4. 라스트 샘플링(Last Sampling)

[0068] 마지막으로, 라스트 샘플링 방법(제4 샘플링 방법으로 명명될 수도 있음)은 가장 낮은 스코어의 데이터를 추출하는 것이다. 탑 샘플링 방법과의 성능 비교를 통해, 스코어링 함수에 의해 발생하는 편향을 알 수 있다. 라스트 샘플링 방법에 의해 구축된 필터링된 데이터 D_f는 수학식 10과 같이 나타낼 수 있다.

[0069] [수학식 10]

[0070] $$D_f = D_{(-score,k)}$$

[0071] 상술한 샘플링 방법들에 대한 성능 평가를 진행한 결과, 탑 샘플링보다는 미들 샘플링이 전체적으로 높은 성능을 보임을 확인하였다. 또한, 대부분의 경우에 메디안 샘플링이 우수한 성능을 보이는 것으로 관찰되었다. 또한, 도 1에는 필터링 방법, 즉 스코어링 함수에 따른 성능 평가 표가 도시되어 있다.

[0072] 도 2는 본 발명의 일 실시예에 따른 병렬 말뭉치 정제 방법을 설명하기 위한 흐름도이다.

[0073] 병렬 말뭉치 정제 방법은 적어도 프로세서(processor) 및/또는 메모리(memory)를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 따라서, 병렬 말뭉치 정제 방법을 이루는 단계들 중 적어도 일부는 병렬 말뭉치 정제 장치로 명명될 수 있는 컴퓨팅의 프로세서의 동작으로 이해될 수 있다. 또한, 컴퓨팅 장치는 PC(Personal Computer), 서버(server) 등을 포함할 수 있다. 또한, 병렬 말뭉치는 NMT(신경망 기반 기계 번역) 모델의 학습에 이용되는 학습 데이터를 의미할 수 있다.

[0074] 우선, 병렬 말뭉치가 확득될 수 있다(S100). 병렬 말뭉치는 각각이 제1 언어로 된 문장과 제2 언어로 된 문장을 포함하는 문장쌍들의 집합을 의미할 수 있다. 병렬 말뭉치는 크롤링 동작을 통해 수집될 수 있다.

[0075] 다음으로, 수집된 병렬 말뭉치에 포함되는 문장쌍들에 대한 스코어링 동작이 수행될 수 있다(S200). 스코어링 동작은 상술한 3가지의 방법(필터링 방법) 중 적어도 하나를 이용하여 수행될 수 있다. 따라서, 필터링 방법별로 문장쌍들에 대한 스코어가 산출될 수 있다.

[0076] 문장쌍들에 대한 스코어가 산출된 후에, 스코어링 함수(또는 필터링 방법이 선택될 수 있다(S300).

[0077] 예시적으로, 학습 대상 모델인 기계 번역 모델이 정해진 경우, 제1 필터링 방법이 선택될 수 있다. 이때, 제1 필터링 방법에 의한 스코어링 함수 중에서 학습 대상 모델에 대하여 가장 높은 성능을 보이는 스코어링 함수가 선택될 수 있다.

[0078] 또한, 문장쌍들에 대한 스코어가 전체적으로 낮은 상황인 경우, 제2 필터링 방법이 선택될 수 있다. 이때, 제2 필터링 방법에 의한 스코어링 함수 중에서 어느 하나가 선택될 수 있다. 이를 위해, 산출된 스코어의 평균값이 산출될 수 있다. 산출된 평균값이 미리 정해진 제1 임계값 보다 작은 경우, 제2 필터링 방법이 선택될 수 있다.

[0079] 또한, 문장쌍들에 대한 스코어가 전체적으로 높은 상황인 경우, 제3 필터링 방법이 선택될 수 있다. 이때, 제3 필터링 방법에 의한 스코어링 함수 중에서 어느 하나가 선택될 수 있다. 이를 위해, 산출된 평균값이 미리 정해진 제2 임계값(제2 임계값은 제1 임계값 보다 큰 임의의 실수) 보다 큰 경우, 제3 필터링 방법이 선택될 수 있다.

[0080] 다음으로, 문장쌍들에 대한 샘플링 동작이 수행될 수 있다(S400). 샘플링은 제2 샘플링 또는 제3 샘플링 방법을 의미할 수 있다. 즉, 수집된 병렬 코퍼스에서 제2 샘플링 또 제3 샘플링 방법을 적용함으로써, 기계 번역의 학습에 사용될 학습 데이터를 구축할 수 있다.

[0081] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨

어 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.

[0082] 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 매체에 저장될 수 있다.

[0083] 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.

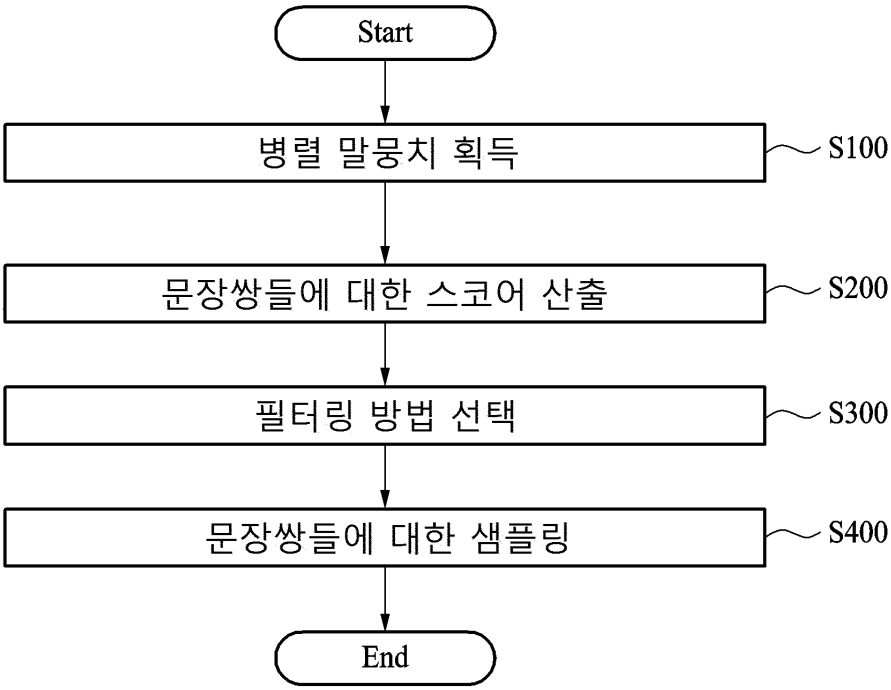
[0084] 본 발명은 도면에 도시된 실시예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시예가 가능하다는 점을 이해할 것이다. 예를 들어, 설명된 기술들이 설명된 방법과 다른 순서로 수행되거나, 및/또는 설명된 시스템, 구조, 장치, 회로 등의 구성 요소들이 설명된 방법과 다른 형태로 결합 또는 조합되거나, 다른 구성 요소 또는 균등물에 의하여 대치되거나 치환되더라도 적절한 결과가 달성될 수 있다. 따라서, 본 발명의 진정한 기술적 보호 범위는 첨부된 등록청구범위의 기술적 사상에 의해 정해져야 할 것이다.

도면

도면1

Test	FLORES								AIhub							
Train	AIhub		CC-Matrix		AIhub		CC-Matrix		AIhub		CC-Matrix		AIhub		CC-Matrix	
	Ko2En	En2Ko	Ko2En	En2Ko	Ko2Ja	Ja2Ko	Ko2Ja	Ja2Ko	Ko2En	En2Ko	Ko2En	En2Ko	Ko2Ja	Ja2Ko	Ko2Ja	Ja2Ko
Random	18.26	17.12	**12.78**	**19.79**	17.56	15.99	**17.59**	**15.96**	**31.39**	**28.81**	**8.40**	**12.69**	**65.50**	**68.62**	**17.99**	**26.35**
TRs	16.49	16.32	6.02	6.41	17.65	16.11	7.41	4.10	22.94	20.81	3.24	3.58	64.53	67.53	10.32	15.16
TRs_dual	14.64	14.41	6.46	10.90	17.83	16.08	8.37	4.90	22.29	20.15	4.25	7.15	64.55	67.50	11.13	15.91
LMs	14.19	12.56	4.26	3.46	17.61	15.86	10.41	5.82	22.52	18.85	2.97	1.42	64.41	66.22	12.41	9.79
LMs_dual	15.74	13.43	6.71	9.05	17.50	15.71	13.20	8.51	28.14	24.92	4.77	5.82	64.53	65.17	14.50	11.15
LASER	**18.94**	**20.78**	5.65	9.25	**17.86**	**16.56**	8.31	5.71	26.35	23.86	3.51	5.02	64.16	67.02	14.42	21.92
LABSE	15.83	16.34	11.20	17.78	17.18	15.70	11.62	8.61	23.97	22.21	8.27	11.25	62.03	64.18	16.47	23.11

도면2


Start
병렬 말뭉치 획득
S100
문장쌍들에 대한 스코어 산출
S200
필터링 방법 선택
S300
문장쌍들에 대한 샘플링
S400
End