(19) 대한민국특허청(KR)
(12) 공개특허공보(A)

(11) 공개번호 10-2022-0033652
(43) 공개일자 2022년03월17일

(51) 국제특허분류(Int. Cl.)
G06F 40/58 (2020.01) *G06F 40/211* (2020.01)
G06F 40/226 (2020.01) *G06F 40/253* (2020.01)
G06N 20/00 (2019.01)
(52) CPC특허분류
G06F 40/58 (2020.01)
G06F 40/211 (2020.01)
(21) 출원번호 10-2020-0115560
(22) 출원일자 2020년09월09일
심사청구일자 2020년09월09일

(71) 출원인
고려대학교 산학협력단
서울특별시 성북구 안암로 145, 고려대학교 (안암동5가)
(72) 발명자
임희석
서울특별시 성북구 안암로 145 자연계캠퍼스애기능생활관 311호 NLP&AI연구실
박찬준
서울특별시 송파구 중대로24, 306동 1303호
(74) 대리인
김홍석

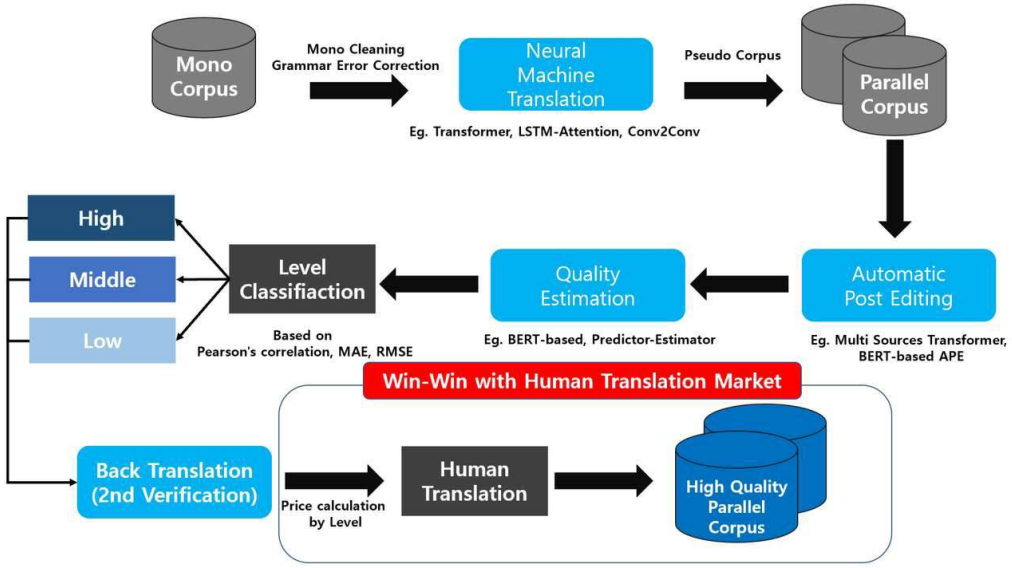
전체 청구항 수 : 총 5 항

(54) 발명의 명칭 **기계 번역의 학습 데이터 구축을 위한 방법**

(57) 요 약

기계 번역 학습 데이터 구축을 위한 방법을 개시한다. 방법은, 원문에 대하여, 단일 말뭉치 클리닝(mono corpus cleaning) 및 문법 교정(grammar error correction)을 수행하는 전처리 단계, 상기 전처리된 단일 말뭉치를 인공 신경망 기계번역 시스템을 이용하여 번역하여 병렬 말뭉치(parallel corpus)를 구축하는 단계, 상기 번역된 병렬 말뭉치를 APE(automatic post editing)하여 번역문을 획득하는 단계, 상기 원문과 상기 번역문을 기반으로, 상기 번역문의 번역 품질을 레벨화하는 단계, 및 상기 번역 품질 결과에 따라 학습 데이터 채택 여부를 결정하는 단계를 포함한다.

대 표 도 - 도1



(52) CPC특허분류
G06F 40/226 (2020.01)
G06F 40/253 (2020.01)
G06N 20/00 (2021.08)

이 발명을 지원한 국가연구개발사업

과제고유번호	1711116144
과제번호	2018-0-01405-003
부처명	과학기술정보통신부
과제관리(전문)기관명	정보통신기획평가원
연구사업명	정보통신방송혁신인재양성(R&D)
연구과제명	Human-inspired 복합지능 원천기술 연구 및 인력양성
기 여 율	1/1
과제수행기관명	고려대학교산학협력단
연구기간	2020.01.01 ~ 2020.12.31

명 세 서

청구범위

청구항 1

적어도 프로세서를 포함하는 컴퓨팅 장치에서 수행되는 기계 번역의 학습 데이터 구축 방법에 있어서,

제1 언어로 작성된 원문을 전처리하는 단계;

인공 신경망 기계번역 모델을 이용하여 상기 원문을 번역하여 번역문을 생성하는 단계;

상기 원문과 상기 번역문을 기반으로, 상기 번역문의 번역 품질을 평가하는 단계; 및

번역 품질 결과에 따라, 상기 원문과 상기 번역문을 학습 데이터에 포함시킬지 여부를 결정하는 단계를 포함하는 기계 번역의 학습 데이터 구축 방법.

청구항 2

제1항에 있어서,

상기 번역 품질을 평가하는 단계에서,

상기 번역문 내 문장 수준의 성능 평가는 피어슨 상관관계(Pearson's correlation), 평균 오차(Mean Average Error, MAE), 및 평균 제곱근 오차(Root Mean Squared Error, RMSE) 중 적어도 하나를 사용하며,

상기 번역문 내 어절 및 구 수준의 성능 평가는 F1-OK 및 F1-BAD의 곱(multiplication)이 사용되는,

기계 번역의 학습 데이터 구축 방법.

청구항 3

제1항에 있어서,

상기 번역 품질을 평가하는 단계는 상기 번역문의 품질을 고-레벨(high level), 중-레벨(middle-level), 및 저-레벨(low-level) 중 어느 하나로 평가하는,

기계 번역의 학습 데이터 구축 방법.

청구항 4

제3항에 있어서,

상기 번역문을 학습 데이터에 포함시킬지 여부를 결정하는 단계는,

상기 번역문의 품질이 고-레벨인 경우, 상기 원문과 상기 번역문을 학습 데이터에 포함시키는,

기계 번역의 학습 데이터 구축 방법.

청구항 5

제1항에 있어서,

상기 전처리하는 단계는 단일 말뭉치 클리닝(mono corpus cleaning) 및 문법 교정(grammar error correction)을 수행하는,

기계 번역의 학습 데이터 구축 방법.

발명의 설명

기 술 분 야

[0001] 본 발명은 기계 번역의 학습 데이터 구축을 위한 방법에 관한 것으로, 보다 상세하게는 기계 번역을 통한 고품

질의 학습 데이터 구축을 위한 방법에 관한 것이다.

배 경 기 술

[0002] 딥 러닝(deep learning)의 등장으로 인공신경망 기계 번역(Neural Machine Translation)이 개발되면서 사람들이 이전보다 만족할 만한 성능의 기계번역기들이 개발되고 있다. 과거 기계번역 연구는 규칙기반 및 통계기반 방식을 이용했으나 최근에는 딥 러닝 기반 방식으로 많은 기술적인 성과를 이루어냈다.

[0003] 특히, GPU의 등장으로 인한 행렬 연산의 병렬 처리화 및 컴퓨팅 파워의 개선, Tensorflow, Pytorch 등의 오픈소스 프레임워크의 등장으로 인한 개발환경의 개선, 웹을 통한 빅데이터 확보 가능, 획기적인 딥 러닝 모델 개발 등의 요인으로 현재 딥 러닝을 이용한 다양한 분야에서 엄청난 성과를 보이고 있으며 기계번역도 마찬가지이다. 그러나 여전히 사람들에게 실질적인 만족감을 부여하기 위해서는 아직 개선되어야 할 사항들이 많다.

[0004] 딥 러닝에서 획기적인 모델 아키텍처를 개발하거나 하이퍼파라미터 최적화를 하는 단계도 중요하지만 무엇보다 중요한 것은 고품질의 학습데이터를 구성하는 일이다. 즉, 딥 러닝 기반 기계번역 시스템의 학습데이터를 구축할 때 양질의 데이터를 확보하는 것이 무엇보다 중요하다. 그러나 양질의 데이터를 구하는 일은 저작권 확보의 문제 및 병렬 말뭉치 구축의 어려움 등을 이유로 쉽지 않은 상황이다.

[0005] 상세하게 설명하면, 인공신경망 기계번역 시스템 관련 데이터 확보 및 구축의 한계는 다음과 같다. 먼저 저작권이 해결된 데이터를 확보하기 어렵다. 이종 언어로 된 병렬 말뭉치를 구하기 어렵다. 더불어 병렬말뭉치의 정제 및 정렬에는 고도의 기술이 필요하며 단일 말뭉치를 원하는 이종 언어로 번역하는 데는 많은 비용과 시간이 필요하다. 기본적인 인공신경망 기계번역의 학습을 위해서는 최소 200만 이상의 병렬 말뭉치가 필요한데 이런 대용량의 인공지능 학습용 말뭉치를 준비하기는 쉽지 않은 실정이다.

[0006] 최근 많은 기업에서 비즈니스를 위하여 도메인 특화 모델을 개발하고 있다. 도메인 특화란 먼저 기업 혹은 학교가 기보유한 데이터로 pre-train 모델을 만들고 도메인특화 데이터로 Fine-tuning을 진행하는 개념으로 기계번역에선 분야에서는 특허번역이 대표적이다. 도메인 특화 모델을 만들게 되면 해당 도메인에서 generic 모델보다 매우 우수한 성능을 보일 수 있는 장점이 존재하나 초기 번역 솔루션 개발을 위한 도메인 데이터 구축을 위해 많은 비용과 시간을 투자해야 한다.

[0007] 또한, Transformer와 같은 좋은 모델이 나오고 구글번역기, 파파고, PNMT 등 좋은 서비스들이 많이 나오고 있는 실정이나 아직까지 품질적 한계가 존재한다. 이로 인한 기계번역 품질 한계로 수요자들의 글로벌 비즈니스 지원 한계가 발생하게 된다. 또한, 인공신경망 기계번역 오픈소스가 활성화되어 전 세계 다수 기업들이 개발 및 상용화를 진행하려 하나, 딥 러닝을 통해 기계번역 솔루션의 성능향상과 고도화를 이루기에는 많은 기술과 데이터 등의 자원들이 부족한 상황이다. 즉, 4차 산업혁명 시대를 맞이하며, 다양한 인공지능 기반 기술·서비스 기업들이 형성되고 관련된 사업들이 증가하고 있으나, 그에 준하는 기술과 데이터 부족에 따른 사업화에 많은 애로사항을 겪고 있다.

[0008] 최근 이러한 문제점을 극복하기 위해 최근 크라우드소싱을 적용한 휴먼번역 시장 확대되고 있다. 크라우드소싱(Crowdsourcing)이라는 단어는 대중 또는 군중이라는 뜻의 'Crowd'와 외부자원활용, 즉 'Outsourcing'의 합성어로 기업의 생산, 서비스 및 문제해결 과정 등에 특정 커뮤니티 또는 불특정 다수의 대중들을 참여토록 하여 효율성을 높이고자 하는 접근방법이다. 즉, 크라우드소싱이라는 집단지성을 통해 데이터를 확보, 정제하여 인공지능 학습용 데이터를 구축한다면 보다 단시간 내 대량의 고품질 데이터 확보가 용이하다. 그러나, 크라우드소싱을 통해서 데이터를 구축하는 것도 어느 정도 한계가 있다. 동일한 문장에 대해 사람마다 다양한 번역 결과가 나올 수 있으며 어떠한 문장을 최적의 데이터(Gold Data)로 정할지도 난해하다.

선행기술문헌

특허문헌

[0009] (특허문헌 0001) 대한민국 공개특허 제2020-0044201호 (2020.04.29. 공개)

(특허문헌 0002) 대한민국 공개특허 제2017-0053527호 (2017.05.16. 공개)

(특허문헌 0003) 대한민국 공개특허 제2019-0097629호 (2019.08.21. 공개)

(특허문헌 0004) 대한민국 공개특허 제2019-0091103호 (2019.08.05. 공개)

발명의 내용

해결하려는 과제

[0010] 본원 발명이 해결하고자 하는 과제는 기계 번역을 통한 고품질의 학습 데이터 구축을 위한 방법에 관한 것이다.

과제의 해결 수단

[0011] 해결하고자 하는 과제를 달성하기 위하여 본 발명의 실시 예들에 따른 기계 번역의 학습 데이터 구축 방법은, 원문에 대하여, 단일 말뭉치 클리닝(mono corpus cleaning) 및 문법 교정(grammar error correction)을 수행하는 전처리 단계, 상기 전처리된 단일 말뭉치를 인공 신경망 기계번역 시스템을 이용하여 번역하여 병렬 말뭉치(parallel corpus)를 구축하는 단계, 상기 번역된 병렬 말뭉치를 APE(automatic post editing)하여 번역문을 획득하는 단계, 상기 원문과 상기 번역문을 기반으로, 상기 번역문의 번역 품질을 레벨화하는 단계, 및 상기 번역 품질 결과에 따라 학습 데이터 채택 여부를 결정하는 단계를 포함한다.

발명의 효과

[0012] 본 발명의 실시 예들에 따르면, 단일 말뭉치만을 이용하여 고품질의 병렬 말뭉치를 구축할 수 있다.

[0013] 또한, 번역문의 번역 품질을 레벨화하여 고-레벨의 번역문은 학습 데이터로 사용되고, 저-레벨의 번역문은 이차 감수를 진행함으로써, 휴먼 번역 감수 가격을 다르게 책정하여 시간을 단축할 수 있으며 감수 작업의 효율을 향상시킬 수 있다. 결론적으로, 시간과 비용의 절약 및 감수 작업의 효율을 향상시켜 고품질 병렬 말뭉치를 확보할 수 있다.

도면의 간단한 설명

[0014] 도 1은 본 발명의 일 실시 예에 따른 기계 번역의 학습 데이터 구축 방법을 설명하기 위한 모식도이다.

발명을 실시하기 위한 구체적인 내용

[0015] 본 발명의 구성 및 효과를 충분히 이해하기 위하여, 첨부한 도면을 참조하여 본 발명의 바람직한 실시 예들을 설명한다. 그러나 본 발명은, 이하에서 개시되는 실시 예들에 한정되는 것이 아니라, 여러 가지 형태로 구현될 수 있고 다양한 변경을 가할 수 있다.

[0016] 또한, 본 발명의 실시 예들에서 사용되는 용어들은 다르게 정의되지 않는 한, 해당 기술 분야에서 통상의 지식을 가진 자에게 통상적으로 알려진 의미로 해석될 수 있다.

[0017] 본 발명에서 구축하는 방향성은 크게 4가지로, 도메인 기반 병렬 말뭉치(Parallel Corpus) 구축, 균형 말뭉치(Balance Corpus) 구축, 병렬 코퍼스 필터링 및 정제, 및 휴먼 번역과의 상생이다. 이하에서 상세하게 설명하기로 한다.

[0018] 도메인 기반 병렬 말뭉치 구축

[0019] 세상에 존재하는 문서에는 여러 전문 도메인이 존재한다. 하지만, 하나의 기계번역 엔진이 모든 분야(도메인)를 모두 정확하게 번역하기는 어렵다. 예로 "trans" 의 경우, 경제 용어로는 "주식거래", 일반용어로 "수송", "번역" 등 10가지 이상으로 다양하게 쓰이기 때문에 각 도메인에 충실한 번역을 위해서는 도메인별 병렬 말뭉치가 필요하다. 도메인 기반으로 데이터를 구축할 시 사용자에게 필요한 데이터를 선별하여 제공해 줄 수 있으며, 관리가 용이하다. 또한, 도메인 기반으로 데이터를 구축할 시 인공신경망 기계번역(NMT) 모델을 만들 때 도메인 특화 모델 제작이 용이하다. 도메인을 기반으로 데이터를 구축하지 않을 경우 데이터 관리가 용이하지 못하여 과적합(Overfitting), 과소적합(Underfitting) 등의 현상이 발생하여 추적이 쉽지 않다. 전문성이 강한 도메인과 포괄적인 도메인의 비중을 달리하여 데이터를 구축하는 것이 양질의 인공신경망 기계번역 시스템(NMT) 모델을 만들 수 있다. 이를 위해 도메인별로 데이터를 구축해야 한다. 도메인은 정치, 경제, 사회, 생활, 문화, 세계, IT, 과학 등으로 나눌 수 있으며 추가적으로 논문, 특허 등도 고려될 수 있다.

[0020]

[0021] 균형 말뭉치 구축

[0022] 균형 말뭉치란 같은 문서 내 정확한 정렬과 번역으로 이루어진 문장들의 집합이다. 대표성이 있는 균형 말뭉치를 작성하기 위해 가장 중요한 점은 다양한 장르를 포괄해야 한다는 점이다. 따라서, 뉴스 말뭉치에서 일반인의 사용 빈도가 높은 사회, 스포츠, 문화, 국제, 과학 분야 등 다양한 도메인 데이터와 도메인 특화를 위한 법률 도메인 데이터를 구축해야 한다. 즉, 하나의 도메인만을 가지고 병렬 말뭉치를 구축하는 것이 아닌 다양한 도메인을 포괄적으로 포함하는 병렬 말뭉치를 구축하는 것이 중요한 요소이다.

[0024] 병렬 코퍼스 필터링 및 정제

[0025] 정제 및 필터링 작업을 거친 말뭉치로 학습을 한 모델이 그렇지 않은 모델보다 BLEU(bilingual evaluation understrudy)점수가 더 높게 나오는 연구가 다수 존재한다. 데이터의 양도 중요하나 대부분의 기업들은 고품질의 데이터를 구축하기 위해 노력 중이다. 인공신경망으로 만든 모델은 양질의 데이터가 성능과 직결되기 때문이다. 병렬 코퍼스 필터링 및 정제를 통해 선별된 말뭉치로 학습을 한 모델이 그렇지 않은 모델보다 더 좋은 BLEU 점수를 보인다. 이는 통계기반 및 규칙기반 방식에서는 데이터의 양이 많을수록 좋았으나 인공신경망 방식에서는 데이터의 양보다는 데이터의 질이 더 중요함을 알 수 있다. 좋지 못한 품질의 데이터는 결국 기계번역 결과값의 오류를 발생시키고 품질하락으로 이어진다. 따라서 고품질의 데이터를 걸러낼 수 있는 병렬 코퍼스 필터링 및 정제에 대한 연구가 이루어져야 할 것이다.

[0027] 휴먼 번역과의 상생

[0028] 데이터의 품질 향상을 위하여 궁극적으로 사람의 손을 거치는 것이 가장 신뢰성 있고 고품질의 데이터가 구축될 수 있다. 그러나 모든 데이터에 대하여 사람의 손을 거쳐서 데이터를 구축하게 된다면 막대한 비용과 시간이 투자되어야 할 것이다. 따라서 어느 정도 컴퓨터가 자동으로 품질을 판단하여 어느 일정 수준 이상의 품질이면 사람의 손을 거치지 않고 일정 수준 미만의 데이터만 사람의 손을 거쳐 검증 및 후처리 작업을 진행하면 좋을 것이다.

[0030] 한편, 병렬 말뭉치를 구축하는 것은 많은 시간과 돈이 들며, 대부분 웹 크롤링(Web Crawling)을 통하여 단일 말뭉치 데이터만 가지고 있다. 이를 위하여 병렬 말뭉치를 만드는 방법을 제시하고자 한다.

[0031] 이하에서는 발명의 일 실시 예에 따른 기계 번역의 학습 데이터 구축 방법을 첨부된 도면과 함께 보다 상세하게 설명하기로 한다. 상기 방법은 적어도 프로세서 및/또는 메모리를 포함하는 컴퓨팅 장치에 의해 수행되는 방법을 의미할 수 있다. 구체적으로 상기 방법은 상기 컴퓨팅 장치의 프로세서에 의해 수행되는 것으로 이해될 수도 있다.

[0032] 도 1은 본 발명의 일 실시 예에 따른 기계 번역의 학습 데이터 구축 방법을 설명하기 위한 모식도이다.

[0033] 도 1을 참조하면, 원문에 대하여, 단일 말뭉치 클리닝(mono corpus cleaning) 및 문법 교정(grammar error correction)하는 전처리 과정을 수행할 수 있다.

[0034] 웹 크로링을 통하여 얻은 데이터들은 문법이 잘못된 경우가 많으며 검증되지 않은 데이터들이기 때문에, 원문의 전처리 과정을 수행하여, 원문의 단일 말뭉치의 질을 높일 수 있다.

[0035] 전처리된 단일 말뭉치를 인공 신경망 기계번역 시스템을 이용하여 번역하여 병렬 말뭉치(parallel corpus)를 구축할 수 있다. 일 실시 예에 따르면, 인공 신경망 기계번역 시스템은 Transformer, LSTM-어텐션, 또는 Conv2Conv 중 하나의 모델로 구현될 수 있다.

[0036] 번역된 병렬 말뭉치를 APE(automatic post editing)하여 번역문을 획득할 수 있다. 일 실시예에 따르면, APE는 multi sources Transformer 또는 BERT 기반 APE 중 하나의 모델로 구현될 수 있다.

[0037] 원문과 번역문을 기반으로, 번역문의 번역 품질을 예측(Quality Estimation)할 수 있다. 번역 품질의 평가는 BERT-기반 또는 Predictor-Estimator 중 하나의 모델로 구현될 수 있다.

[0038] 번역 품질의 성능을 점수로 나타낼 수 있다. 일 실시예에 따르면, 번역문 내 문장 수준의 성능 평가는 피어슨

상관관계(Pearson's correlation), 평균 오차(Mean Average Error, MAE), 및 평균 제곱근 오차(Root Mean Squared Error, RMSE) 중 적어도 하나를 사용하며, 번역문 내 어절 및 구 수준의 성능 평가는 F1-OK 및 F1-BAD의 곱(multiplication)이 사용할 수 있다.

[0039] 점수를 척도로 이용하여 휴먼번역을 위한 감수 레벨을 정하게 될 수 있다.

[0040] 일 실시 예에 따르면, 3단계(High, Middle, Low)의 감수 레벨로 분류하여 고-레벨의 번역문은 학습 데이터로 사용하고, 중-레벨 및/또는 저-레벨의 번역문은 이차 감수를 진행할 수 있다.

[0041] 중-레벨 및/또는 저-레벨의 번역문을 이차 감수 시, 번역문을 휴먼 번역을 통해 재번역할 것인지를 결정할 수 있다. 이차 감수는 백 번역(back translation)을 적용하여 이차 감수를 진행할 수 있다.

[0042] 즉, 이차 감수 후 사용자는 번역문을 말뭉치 데이터로 사용할지 아니면 저렴한 가격으로 번역 감수를 맡길지 결정하게 된다. 즉, 이차 감수 후 사용자는 번역문을 말뭉치 데이터로 사용할지 아니면 저렴한 가격으로 번역 감수를 맡길지 결정하게 된다. 이를 통해 전문 번역가 이차 감수 과정을 통하여 양질의 인공지능 학습용 데이터를 구축하게 될 것으로 기대된다.

[0043] 본 발명의 일 실시 예에 따른 기계 번역의 학습 데이터 구축을 위한 방법을 이용하면, 단일 말뭉치만을 이용하여 고품질의 병렬 말뭉치를 구축할 수 있다. 또한, 번역문의 번역 품질을 레벨화하여 고-레벨의 번역문은 학습 데이터로 사용되고, 저-레벨의 번역문은 이차 감수를 진행함으로써, 휴먼 번역 감수 가격을 다르게 책정하여 시간을 단축할 수 있으며 감수 작업의 효율을 향상시킬 수 있다. 결론적으로, 시간과 비용의 절약 및 감수 작업의 효율을 향상시켜 고품질 병렬 말뭉치를 확보할 수 있다.

[0045] 상술한 기계 번역의 학습 데이터 구축을 위한 방법(즉, 학습 데이터 생성 방법)을 다시 설명하면 다음과 같다. 학습 데이터 생성 방법은 적어도 프로세서 및/또는 메모리를 포함하는 컴퓨팅 장치에 의해 수행될 수 있다. 즉, 학습 데이터 생성 방법은 컴퓨팅 장치에 포함되는 프로세서의 동작을 의미할 수도 있다.

[0046] 우선, 웹 크롤링 등을 통해 미리 저장된 원문을 대상으로 전처리 동작이 수행될 수 있다. 전처리 동작을 통해 원문의 단일 말뭉치의 질을 높일 수 있다. 이때, 원문은 제1 언어(예컨대, 한글 또는 영문)로 작성된 문장을 의미할 수 있으며, 전처리 동작은 단일 말뭉치 클리닝 및 문법 교정을 의미할 수 있다.

[0047] 다음으로, 소정의 인공 신경망 기계번역 모델을 이용하여 상기 원문 또는 전처리된 원문을 번역하여 상기 원문에 대응하는 번역문을 생성함으로써 병렬 말뭉치를 생성할 수 있다. 번역문은 제2 언어(예컨대, 영문 또는 한글)로 작성된 문장을 의미할 수 있으며, 기계번역 모델은 트랜스포머(Transformer), LSTM-어텐션, 및/또는 Conv2Vonv 중 어느 하나를 의미할 수 있다.

[0048] 병렬 말뭉치는 기계번역 사후 교정 모델, 예컨대 multi sources Transformer, BERT-based APE 등과 같은 모델을 이용하여 APE(automatic post editing)함으로써, 번역문의 품질을 보다 향상시킬 수 있다.

[0049] 이후, 원문과 번역문을 기반으로, 번역문의 번역 품질이 평가될 수 있다. 번역 품질의 평가는 BERT-기반 또는 Predictor-Estimation 중 어느 하나를 이용하여 수행될 수 있다.

[0050] 번역문의 품질 평가 결과에 기초하여, 원문과 번역문을 학습 데이터에 포함시킬지 여부가 결정된다. 예컨대, 원문에 대한 번역문이 미리 정해진 품질 이상의 품질을 갖는 것으로 평가된다면, 원문과 번역문은 학습 데이터로 활용될 수 있다. 구체적으로, 3 단계(고-레벨, 중-레벨, 및 저-레벨)로 품질 평가 결과가 산출되는 경우, 고-레벨(또는 중-레벨과 고-레벨) 평가되는 경우에 한해서 학습 데이터로 이용될 수 있다.

[0051] 이후, 전 단계에서 원문과 번역문이 학습 데이터로 이용되지 않는 경우, 번역문문을 기준으로 백 번역이 수행될 수 있으나, 이 과정은 생략될 수 있다.

[0052] 다음으로, 학습 데이터로 이용되지 않는 원문과 번역문은 전문 번역가에 의한 이차 감수 과정을 통하여 양질의 학습 데이터로 포함될 수 있다. 즉, 번역문은 전문 번역가에 의해 수정 및/또는 보완되어 보다 높은 품질의 번역문이 생성될 수 있다. 이를 위해, 컴퓨팅 장치는 2차 감수 대상인 원문과 번역문을 사용자(즉, 전문 번역가)의 단말기로 송신하고, 감수가 완료된 원문 및/또는 번역문을 사용자의 단말로부터 수신할 수 있다. 사용자 단말기로부터 수신된 원문과 번역문은 학습 데이터로 포함될 수 있다.

[0054] 이상에서 설명된 장치는 하드웨어 구성 요소, 소프트웨어 구성 요소, 및/또는 하드웨어 구성 요소 및 소프트웨어 구성 요소의 집합으로 구현될 수 있다. 예를 들어, 실시 예들에서 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.

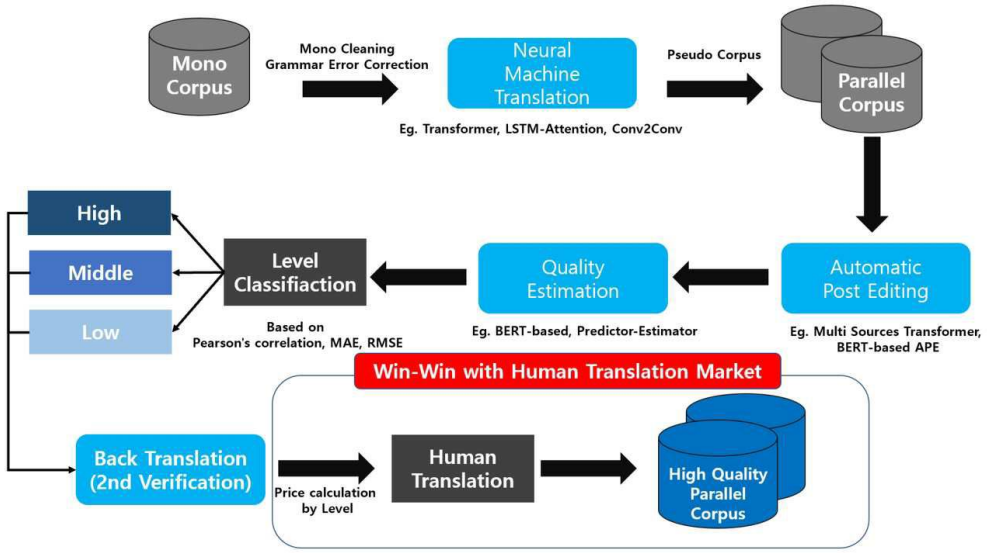
[0055] 소프트웨어는 컴퓨터 프로그램(Computer Program), 코드(Code), 명령(Instruction), 또는 이들 중 하나 이상의 조합을 포함할 수 있으며, 원하는 대로 동작하도록 처리 장치를 구성하거나 독립적으로 또는 결합적으로(Collectively) 처리 장치를 명령할 수 있다. 소프트웨어 및/또는 데이터는, 처리 장치에 의하여 해석되거나 처리 장치에 명령 또는 데이터를 제공하기 위하여, 어떤 유형의 기계, 구성 요소(Component), 물리적 장치, 가상 장치(Virtual Equipment), 컴퓨터 저장 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.

[0056] 실시 예에 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 좋ㅂ하여 포함할 수 있다. 상기 매체에 기록되는 프로그램 명령은 실시 예를 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 실시 예의 동작을 수행하기 위해 하나 이상의 소프트웨어 모듈로서 작동하도록 구성될 수 있으며, 그 역도 마찬가지이다.

[0057] 본 발명은 도면에 도시된 실시 예를 참고로 설명되었으나 이는 예시적인 것에 불과하며, 본 기술 분야의 통상의 지식을 가진 자라면 이로부터 다양한 변형 및 균등한 타 실시 예가 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


Mono Corpus
Mono Cleaning
Grammar Error Correction
Neural Machine Translation
Eg. Transformer, LSTM-Attention, Conv2Conv
Pseudo Corpus
Parallel Corpus
Automatic Post Editing
Eg. Multi Sources Transformer, BERT-based APE
Quality Estimation
Eg. BERT-based, Predictor-Estimator
Level Classifiaction
Based on Pearson's correlation, MAE, RMSE
High
Middle
Low
Win-Win with Human Translation Market
Back Translation (2nd Verification)
Price calculation by Level
Human Translation
High Quality Parallel Corpus