번역 품질 예측을 위한 HTER 분포 평준화 기반 인조 번역 품질 말뭉치 구축 방법

박준수[O*], 이원기[*], 신재훈[*], 한효정[+], 이종혁[*]

포항공과대학교 컴퓨터공학과[*], 삼성 리서치[+]
{jspak3, wklee, jaehun.shin}@postech.ac.kr, h.j.han@samsung.com, jhlee@postech.ac.kr

Construction of an Artificial Training Corpus for The Quality Estimation Task based on HTER Distribution Equalization

Junsu Park[O*], WonKee Lee[*], Jaehun Shin[*], H. Jeung Han[+], Jong-hyeok Lee[*]
Dept. of Computer Science and Engineering, POSTECH[*]; Samsung Research[+]

요 약

번역 품질 예측은 기계번역 시스템이 생성한 번역문의 품질을 정답 번역문을 참고하지 않고 예측하는 과정으로, 번역문의 사후 교정을 위한 번역 오류 검출의 역할을 담당하는 중요한 연구이다. 본 논문은 문장 수준의 번역 품질 예측 문제를 HTER 구간의 분류 문제로 간주하여, 번역 품질 말뭉치의 HTER 분포 불균형으로 인한 성능 제약을 완화하기 위해 인조 사후 교정 말뭉치를 이용하는 방법을 제안하였다. 결과적으로 HTER 분포를 균등하게 조정한 학습 말뭉치가 그렇지 않은 쪽에 비해 번역 품질 예측에 더 효과적인 것을 보였다.

주제어: 기계 번역, 번역 품질 예측, 이단 구조 번역 품질 예측 모델

1. 서론

번역 품질 예측(Quality Estimation)이란, 기계번역 시스템이 생성한 번역문(Translated sentence)의 품질(Quality)을 정답 번역문(Reference translation)을 참고하지 않고 예측하는 과정이다[1, 2]. 기계 번역은 현재 다양한 곳에서 널리 이용되는 기술이지만, 여전히 기계 번역 시스템이 생성한 번역문에는 번역 오류가 발생한다. 이런 번역 오류는 정답 번역문과의 비교를 통해서만 파악이 가능하지만, 실제로 기계 번역 시스템으로 생성되는 모든 번역문에 대한 정답을 만들어내는 것은 불가능하다. 따라서, 정답이 주어지지 않은 환경에서 기계 번역문의 품질을 예측하는 것은 중요한 의미를 가지며, 이후 번역문을 이용하여 교정된 문장을 생성하는 사후 교정(Post-editing)에도 유용하게 활용될 수 있다.

번역 품질 예측은 모델의 학습을 위한 학습 말뭉치(Training Corpus)로서 번역 품질 말뭉치(Quality Estimation Corpus)를 필요로 하며, 번역 품질 말뭉치는 기본적으로 원시 문장과 번역문, 그리고 해당 번역문에 대한 번역 품질 정보로 이루어진다. 이 번역 품질 정보는 기계가 예측할 레이블(Label) 값으로, 번역문 품질 예측 분야는 예측할 레이블 값에 따라 문장, 구, 단어 수준의 품질 예측 문제로 세분화될 수 있다. 문장 수준의 경우 번역문을 사람이 직접 수정한 정답 번역문(Human-targeted reference)을 기준으로 계산되는 TER(Translation Edit Rate)이 대상 레이블이 되며, 이 때 정답이 사람이 작성한 문장이기에 HTER이라고 불린다. 단어/구 수준의 경우 각 단위가 올바른 번역인지에 따라 OK/BAD 태그가 부여된다. 이러한 번역 품질 말뭉치를 구축하기 위해서는 사람이 정답 번역문을 작성해야 하는 어려움이 있다.

이러한 배경에서, 번역문 사후 교정 연구에서 널리 사용되는 인조 말뭉치(Artificial Corpus)[3, 4]를 번역 품질 예측 연구에 적합한 형태로 변환하여 적용시키는 방법이 제안되었다[5]. 그러나, 이 인조 말뭉치의 정답 번역문은 사람이 작성한 것이 아니기 때문에 적절한 방법을 통해 기존의 번역 품질 말뭉치와 유사한 특성을 가지는 문장만을 추출하여 학습에 사용할 경우 더 나은 성능을 보일 것으로 기대하였다.

본 논문에서는 이러한 점을 고려하여, 인조 사후 교정 말뭉치로 번역 품질 말뭉치를 구축하는 과정에 정답 번역문과 관련된 특성을 이용하는 문장 선별 방법을 적용하고, 실험을 통해 제안한 방법이 문장 수준에서 번역 품질 예측 성능에 미치는 효과를 확인해 보았다.

2. 관련 연구

2.1 번역문 자동 사후 교정을 위한 인조 말뭉치

번역문 자동 사후 교정 연구 분야의 학습 말뭉치는 원시 문장(source sentence)과 이를 임의의 번역 시스템으로 번역한 번역문, 그리고 그 번역문에 대한 교정문(post-edited sentence)으로 구성되어 있다. 이 때 교정문은 사람이 직접 기계 번역문을 최소 수정을 가해 교정된 문장[7]이기 때문에 말뭉치 구축에 많은 비용이 들고, 실제로 존재하는 학습 말뭉치의 양은 모델을 학습하기에 매우 적은 것이 현실이다. 더욱이, 신경망 기반의 시퀀스

변환 모델[8, 9]이 번역문 자동 사후 교정 분야에 적용되면서 대량의 학습 말뭉치에 대한 필요성이 커지게 되었다[10].

이에 따라서 번역문 자동 사후 교정 분야의 학습 말뭉치로서 대량의 단일 언어 말뭉치(Monolingual corpus)를 기반으로 생성한 인조 말뭉치가 제안되었다[3, 4]. 해당 인조말뭉치는 단일 언어 말뭉치에서 '잘 생성된(well-formed)' 문장을 골라내어 교정문으로 삼은 뒤, 이것을 임의의 번역 시스템으로 번역하여 생성된 문장을 원시 문장으로 삼았다. 그리고 다시 이 원시 문장에 대한 번역 시스템의 결과를 목표 언어의 번역문으로서 구성하는, 왕복 번역(Round-trip translation) 방식으로 구축되었다. 이렇게 생성된 문장의 트리플렛(Triplet) – (원시 문장, 번역문, 교정문)들은 번역문과 교정문 사이의 TER를 바탕으로 적절히 선택되어 최종 인조 말뭉치를 구성한다.

신경망 기반의 번역문 자동 사후 교정 학습 시스템에서 해당 인조말뭉치를 사전 학습 과정에 사용함으로서 데이터 부족 문제를 완화하여 자동 사후 교정 성능의 유의미한 향상을 보였기에[3, 4], 이후에 제안된 번역문 자동 사후 교정 연구에서는 기본적으로 해당 인조 말뭉치들을 활용하고 있다[10, 11].

2.2 사후 교정 말뭉치를 이용한 번역 품질 말뭉치 확장

번역 품질 말뭉치는 각 번역문에 대해 적절한 정답 번역문이 있다고 가정하고, 해당 정답 번역문과의 비교로 생성된 TER과 태그 정보로 구성되어 있다. 이는 번역문에 대해 적절한 정답 번역문을 사용한다는 점에서 번역문 자동 사후 교정과 동일한 가정이 반영되어 있으며, 사후 교정 말뭉치는 원시 문장, 번역문, 교정문으로 구성되어 있어 이를 쉽게 번역 품질 말뭉치로 변환할 수 있다. 이러한 가정을 바탕으로, 번역문 자동 사후 교정을 위해 구축된 대량의 인조 말뭉치를 변환하여 필터링하는 방식으로 번역 품질 말뭉치를 확장하는 방법이 등장하였다[5].

2.3 Predictor-Estimator 구조의 번역 품질 예측 모델

Predictor-Estimator 구조는 2017년에 제안된 신경망 기반 번역 품질 예측 구조로, 병렬 말뭉치로 학습되는 이중언어(bilingual) 및 양방향(bidirectional)에 기반한 단어 추정 모델(word prediction model)인 Predictor와 번역 품질 말뭉치로 학습되는 양방향 순환 신경망(BiRNN) 기반의 품질 예측 모델인 Estimator로 구성된다[6]. 그림 1은 Predictor-Estimator의 기본 구조를 나타낸 것이다.

Predictor는 주의 기반(Attention-based)의 신경망 모델을 번역 품질 예측에 적합하게 변형한 것이다. 번역문의 각 단어에 대한 특징 벡터(Quality Estimation Feature Vector, QEFV)를 생성하며, 이를 Estimator에 전달하여 Estimator는 Predictor에서 생성한 특징 벡터를 바탕으로 적절한 품질 예측 점수를 생성한다. 이 모델은 2018년 번역 품질 예측 분야에서 가장 뛰어난 성능을 보였으며[12], 현재도 이 모델을 기반으로 한 번역 품질 예측 모델들이 제안되고 있다[13].

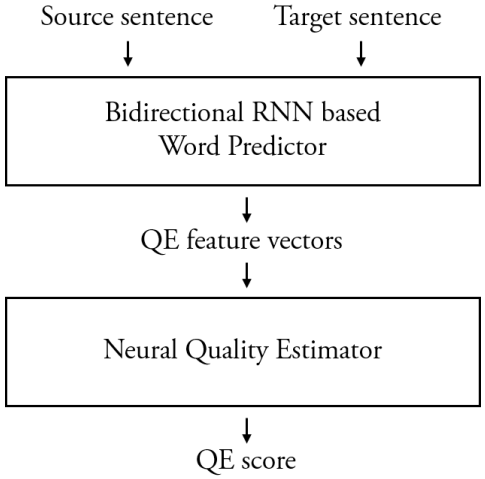

그림 1. Predictor-Estimator의 기본 구조

3. 연구 내용

문장 수준의 번역 품질 예측은 원시 문장과 번역문을 입력으로 하여 품질을 점수로 예측하는 회귀 문제(Regression)로 정의되지만, 이 점수는 계산 상 문장의 길이로 나누어지는 HTER 기반이기 때문에 불연속적인 특징을 지닌다. 따라서 이를 일정 구간 별로 나누어 클래스로 구성하는 것으로 분류 문제로 정의할 수 있다.

분류 문제에서는 클래스 별로 최소한의 학습 데이터가 필요하며, 클래스 별 학습 데이터의 비중이 편향되어 있는 경우 전체적인 성능에 악영향을 미칠 수 있으므로, 분포가 균일하도록 데이터를 구성해야 한다[14]. 하지만, 번역 품질 말뭉치의 경우 0.1 단위의 HTER 분포가 그림 2와 같이 크게 불균형한 분포를 가지고 있으며, 이에 대한 최근 연구에서는 번역 품질 말뭉치의 HTER 분포 불균등을 완화하는 것이 품질 예측 성능에 유의미한 영향을 주는 것이 보고되었다[6].

이런 관점에서, 우리는 인조 사후 교정 말뭉치를 그대로 번역 품질 말뭉치로 변환하는 기존의 방법[5]에 더해, 번역 품질 예측 문제에서 가정하는 정답 번역문과 관련된 특성을 고려하여 말뭉치의 분포를 조정하는 과정을 추가하는 것으로 번역 품질 예측 시스템의 성능을 향상시킬 수 있을 것이라고 기대하였다. 이에 우리는 아래와 같은 가정을 바탕으로 말뭉치 정제 과정을 구성하였다.

I. 인조 말뭉치의 정답 번역문은 번역문을 최소 수정하여 교정된 문장이 아닌 점을 고려하여, 사람이 직접 번역을 수정한 학습 말뭉치와 유사한 특성을 지닌 문장을 선별하는 것으로 이상점(Outlier)를 학습하지 않도록 유도한다.

II. 학습의 목표가 되는 HTER의 구간을 나누고, 이를 기준으로 인조 데이터의 TER을 구간별로 나누어 각 구간이 균등한 분포를 가지도록 문장을 선별하는 것으로 학습과정에서 각 구간에 대해 동일한 비율의 데이터를 이용하도록 유도한다.

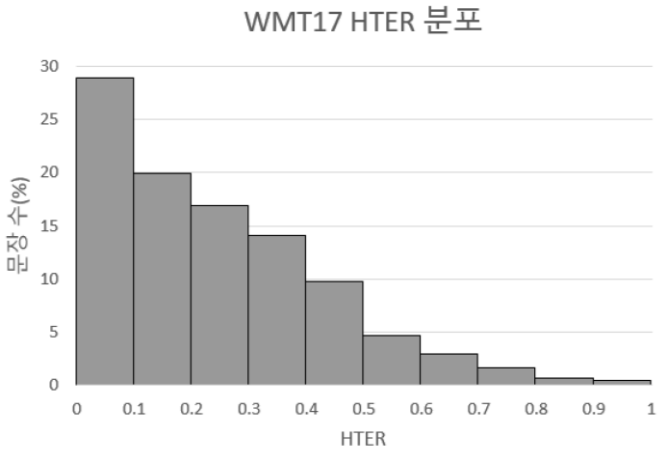

그림 2. WMT17 영어-독일어 학습 말뭉치 HTER 분포

본 논문에서는 I에서 가정한 학습 말뭉치와 유사한 특성을 정의하기 위하여 두 가지 방법을 시도하였다. 첫 번째 방법은 언어 모델에 기존의 학습 데이터를 학습한 뒤, 학습된 모델에 인조 말뭉치의 번역문을 통과시켜 얻는 점수를 기준으로 유사도를 판별한 방법이다. 언어 모델을 이용해 기존의 학습 말뭉치의 특질을 학습하여, 이를 통한 유사도 판별이 가능할 것이라고 기대하였다. 두 번째 방법은 언어 모델 점수를 포함한 특징 벡터를 생성하는 것이다. 번역 품질 예측을 수행할 때, 언어 모델이 학습한 특성만으로는 번역문에서 수정할 단어를 예측하는 것이 힘들 것이라고 예상하였다. 그렇기 때문에 번역문과 교정문 사이의 최소 수정 시 실제로 수행한 삽입, 삭제, 이동 작업 횟수와 TER 값, 학습 말뭉치와의 단어 차이 등의 특성 등이 유사한 문장을 선별하기 위하여 이를 고려한 11차원의 특성 벡터를 생성하였다. 벡터의 각 차원은 다음과 같다.

- 문장 길이
- 최소 수정 시 삽입 횟수
- 최소 수정 시 삭제 횟수
- 최소 수정 시 이동 횟수
- 삽입 비율(삽입 횟수/문장 길이)
- 삭제 비율(삭제 횟수/문장 길이)
- 이동 비율(이동 횟수/문장 길이)
- 미지 단어(Unknown word) 개수
- 미지 단어 비율(미지 단어 개수/문장 길이)
- 언어 모델 점수
- TER

II에서 가정한 HTER의 구간은 0.1 단위로 나누어 오류가 없는(HTER이 0인) 클래스와 완전 수정을 가한(HTER이 1 이상인) 클래스를 포함하여 12개 클래스로 분류하였다. 실험에 사용한 학습 말뭉치의 평균 문장 길이는 17.5, 인조 말뭉치의 평균 문장 길이는 15.8로, 0.05 이하의 값으로 구간을 나누는 것은 의미가 없으리라고 예측하였기 때문이다. HTER 분포에 따른 학습 성능 차이를 실험하기 위하여 학습 말뭉치와 유사한 분포를 갖는 인조 말뭉치(Skewed)와, 모든 HTER 구간이 동일한 인조 말뭉치(Balanced)를 각각 실험하였다.

표 1. HTER 분포에 따른 인조 말뭉치의 분포

HTER(0~1) District	Skewed	Balanced
0.0	15.51%	8.33%
0.0~0.1	13.36%	8.33%
0.1~0.2	19.92%	8.33%
0.2~0.3	16.91%	8.33%
0.3~0.4	14.10%	8.33%
0.4~0.5	9.72%	8.33%
0.5~0.6	4.67%	8.33%
0.6~0.7	3.00%	8.33%
0.7~0.8	1.63%	8.33%
0.8~0.9	0.69%	8.33%
0.9~1.0	0.49%	8.33%
1.0	0.00%	8.33%

4. 실 험

4.1 실험 환경

본 실험에 사용된 모델은 OpenKiwi[15]에 구현된 Predictor-Estimator 모델이며, 모델에 대한 변수 설정 값은 다음과 같다. Predictor 모델에서 각 언어별 처리 언어 개수(Vocabulary size) 45,000개, 단어 임베딩(Word Embedding)의 크기 200, Bi-RNN에서의 hidden state의 크기 400을 사용하였고, 두 번째 단계 모델의 Bi-RNN에서의 hidden state의 크기 125를 사용하였다.

실험에 사용된 1단계 모델의 학습 말뭉치는 OpenKiwi에서 제공한 3,396,364 문장쌍의 영어-독일어 말뭉치를 사용하였다. 2단계 모델의 학습 말뭉치의 경우 23,000개 문장의 WMT17 Quality Estimation의 영어-독일어 학습 말뭉치[17]를 사용하였다. 선별에 사용할 인조 말뭉치는 4,334,934개 문장을 가진 말뭉치[3]를 사용하였으며, 인조 말뭉치는 HTER 분포 구간별로 선별하여 총 69,000개 문장의 말뭉치로 선별되어 학습되었다. 학습 말뭉치는 번역 품질이 높은 (HTER값이 0에 가까운) 구역에 있는 문장이 많이 존재하게 되는 경향(심한 편향성)을 보이고 있으며, 인조 말뭉치를 표 1과 같은 분포를 가지는 Skewed 말뭉치와 모든 구간에서 같은 분포를 가지는 Balanced 말뭉치로 선별하였다. 각 말뭉치를 선별할 때 유사도에 대한 가설을 확인하기 위하여 다음과 같이 다른 방법으로 각각 선별하였다.

i. 유사도를 고려하지 않고 무작위(Random-sampling)

표 2. 각 말뭉치 별 최종 성능(Pearson's Correlation)

Filtering Method	Distribution	
	Skewed	Balanced
All	0.40943	
Naïve1	0.62654	0.61655
Naïve2	0.61378	0.63897
KenLM	0.55888	0.57203
Feature	0.56139	0.58798

로 뽑은 말뭉치(Naïve).

ii. 언어 모델 점수를 기반으로 가장 점수가 높은 상위 6,9000개의 샘플을 뽑은 말뭉치(KenLM).

iii. 3에서 가정한 학습 말뭉치와 인조 말뭉치의 특성 벡터 사이의 코사인 유사도(Cosine similarity)를 비교하여 유사도가 가장 높은 6,9000개의 샘플을 뽑은 말뭉치(Feature).

이 중 Random-sampling을 거친 인조 말뭉치(i)의 경우 무작위의 특성상 서로 다른 seed를 사용하여 서로 다른 말뭉치 두 개를 선별하여 정확한 실험을 기대하였다. 또한 ii에서 사용한 언어 모델로서 KenLM[16]을 사용하였다.

4.2 실험 결과 및 분석

본 실험은 문장 수준에 대한 번역 품질 예측을 수행하였으며, 성능 평가는 WMT의 번역 품질 예측 성능평가 캠페인에서 공식적으로 사용되는 성능 평가 척도(Pearson's correlation)를 사용하였다[13]. 각 실험은 학습된 1단계 모델을 기반으로, 각 인조 말뭉치로 2단계 모델을 사전 학습한 뒤 WMT17 학습 말뭉치로 Fine-tuning을 거쳤다.

표 2는 선별된 인조 말뭉치로 문장 수준의 Predictor-Estimator 모델을 학습하여, 문장 수준의 영어-독일어 번역 품질 예측을 수행한 결과이다. 본 연구에서 수행한 번역 품질 말뭉치 선별이 잘 수행되었는지 확인하기 위해, 4.1에서 선별한 각 말뭉치와 기존 전체 인조 말뭉치(All)를 학습하여 WMT17 영어-독일어 평가 말뭉치에 대하여 품질 예측을 수행하였다.

선별을 거치지 않은 모든 인조 말뭉치를 사용하여 학습한 번역 품질 예측 성능(All)은 상대적으로 적은 양의 말뭉치를 선별한 다른 말뭉치에 비하여 크게 낮았다. 이는 1단계 모델에서 이미 대량의 노이즈가 적은(Clean) 말뭉치로 학습한 다음 이를 기반으로 2단계 모델을 학습하는 Predictor-Estimator의 특성 상, 노이즈가 다량 포함된 인조 말뭉치를 전부 이용하는 경우 과소적합(Underfitting)이 일어나기 때문으로 보인다.

선별을 거치지 않은 인조 말뭉치를 제외한 선별된 말뭉치의 경우, HTER 분포를 학습 데이터와 동일하게 설정한 Skewed 말뭉치보다 Balanced 말뭉치의 성능이 일반적으로 더 좋았다. Skewed 말뭉치의 번역 품질 예측 성능에 비해 Balanced 말뭉치의 경우 각각 -1.59%, 4.10%, 2.35%, 4.74%의 성능 향상을 기록하였다. 이는 기존 학습 말뭉치에서 상대적으로 적은 문장을 학습할 수 밖에 없는 기존 분포에서보다 HTER 분포를 균일하게 맞춘 말뭉치에서 오류율이 높은 문장에 대해서도 견고한 품질 예측이 가능하기 때문이라고 예측된다.

선별 방법의 경우, 동일한 방법으로 무작위로 문장을 추출한 Naïve1, Naïve2 말뭉치가 가장 성능이 좋았다. 시드 값에 따라서 말뭉치의 성능은 차이가 있었지만, 평균적인 성능은 실험 중에서 가장 높았으며, 특성 벡터를 이용한 Feature 말뭉치가 그 다음으로 높았다. 이는 KenLM 말뭉치와 Feature 말뭉치의 평균 문장 길이가 10.6단어로 Naïve 말뭉치의 평균 문장 길이인 15.2단어에 비해 크게 작은 수치인 반면, 이 때문에 평균 문장 길이가 17.8단어로 상대적으로 긴 문장들을 제대로 학습하지 못했기 때문으로 보인다. 각 단어의 출현 확률의 곱으로 계산되는 KenLM 점수 알고리즘의 특성 상, 적은 길이의 문장이 상대적으로 높은 점수를 기록하여 짧은 문장들 위주로 선별되었던 문제점이 있으며, 이는 점수 계산을 길이에 의존하지 않는 언어 모델을 사용함으로서 개선 가능하다고 보여진다.

언어 모델 점수만을 사용하여 품질 예측을 수행한 KenLM 말뭉치에 비하여 Feature 말뭉치의 품질 예측 성능이 더 좋았는데, 이는 사용한 특성 벡터가 품질 예측의 성능에 유의미한 효과를 내고 있다고 분석된다.

5. 결 론

본 논문에서는 HTER 분포의 평준화를 이용하여 번역 품질 예측 말뭉치를 구축해 보았다. 이 과정에서 영어-독일어 번역문 자동 사후 교정 인조 말뭉치를 번역 번역 품질 확장 말뭉치로 변환할 때 시도할 수 있는 방법을 제시하였으며, 말뭉치의 HTER 분포 불균형을 해소한 인조 말뭉치를 구축하였다. 문장 수준에서의 번역 품질 예측에서 HTER 분포가 균일한 말뭉치를 학습하였을 때, 그렇지 않은 말뭉치를 학습할 때보다 좋은 품질 예측 성능을 보임을 확인할 수 있었다.

향후 이러한 사실을 바탕으로, 이후 연구에서는 번역문 자동 사후 교정 인조 말뭉치로부터 번역 품질 예측 문제에 더 적합한 인조 말뭉치 선별에 대해 연구를 진행할 계획이다.

사사

본 연구는 Samsung Research의 지원을 받아 수행한 연구 과제임.

참고문헌

[1] John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing, "Confidence estimation for machine translation," Proc. of the 20th international conference on Computational Linguistics, 2004.

[2] Lucia Specia, Marco Turchi, Nicola Cancedda,

MarcDymetman, and Nello Cristianini, "Estimating the sentence-level quality of machine translations systems," 13th Conference of the European Association for Machine Translation, pp. 28-37, 2009.
[3] Junczys-Dowmunt, Marcin, and Roman Grundkiewicz. "Log-linear Combinations of Monolingual and Bilingual Neural Machine Translation Models for Automatic Post-Editing." Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. 2016.
[4] Negri, Matteo, et al. "ESCAPE: a Large-scale Synthetic Corpus for Automatic Post-Editing." Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018). 2018.
[5] Wang, Jiayi, et al. "Alibaba submission for WMT18 quality estimation task." Proceedings of the Third Conference on Machine Translation: Shared Task Papers. 2018.
[6] Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, Seung-Hoon Na. "Predictor-Estimator: Neural Quality Estimation Based on Target Word Prediction for Machine Translation." ACM Transactions on Asian and Low-Resource Language Information Processing, 17(1). 2017.
[7] Turchi, Marco; Chatterjee, Rajen and Negri, Matteo. "WMT16 APE Shared Task Data" LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University: http://hdl.handle.net/11372/LRT-1632, 2016.
[8] LIBOVICKÝ, Jindřich, et al. "CUNI System for WMT16 Automatic Post-Editing and Multimodal Translation Tasks." Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers. p. 646-654, 2016.
[9] CHATTERJEE, Rajen, et al. "Multi-source Neural Automatic Post-Editing: FBK's participation in the WMT 2017 APE shared task." Proceedings of the Second Conference on Machine Translation. p. 630-638, 2017.
[10] Chatterjee, Rajen, et al. "Findings of the WMT 2018 Shared Task on Automatic Post-Editing." Third Conference on Machine Translation. The Association for Computational Linguistics, pp. 723-738, 2018.
[11] Ondˇrej, Bojar, et al. "Findings of the 2017 conference on machine translation (wmt17)." Second Conference on Machine Translation. The Association for Computational Linguistics, 2017.
[12] Lucia, Specia, et al. "Findings of the WMT 2018 Shared Task on Quality Estimation." Third Conference on Machine Translation. The Association for Computational Linguistics, pp. 702-722, 2018.
[13] Erick, Fonseca, et al. "Findings of the WMT 2019 Shared Tasks on Quality Estimation." Fourth Conference on Machine Translation. The Association for Computational Linguistics, vol. 3, pp. 1-12, 2019.
[14] Chawla N.V. "Data Mining for Imbalanced Dataset: An Overview." Data Mining and Knowledge Discovery Handbook. Springer, Boston, Ma, 2009.
[15] Fabio, Jonay, et al. "OpenKiwi: An Open Source Framework for Quality Estimation." Annual meeting of the Association for Computational Linguistics, 2019.
[16] Kenneth Heafield. "KenLM: Faster and Smaller Language Model Queries." Proceedings of the Sixth Workshop on Statistical Machine Translation, 2011.
[17] Specia, Lucia and Logacheva, Varvara, "WMT17 Quality Estimation Shared Task Training and Development Data", LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11372/LRT-1974, 2017