creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

기계번역학습 데이터셋 품질 평가를 위한 시각적 분석 시스템

A Visual Analytics System for evaluating dataset of Neural Machine Translation

2023 년 2 월

서울대학교 대학원

컴퓨터공학부

박세범

기계번역학습 데이터셋 품질 평가를 위한 시각적 분석 시스템

지도 교수 : 서 진 욱

이 논문을 공학석사 학위논문으로 제출함
2022년 12월

서울대학교 대학원
컴퓨터공학부
박세범

박세범의 공학석사 학위논문을 인준함
2022 년 12 월

위 원 장 _____ 신영길 _____ (인)

부위원장 _____ 서진욱 _____ (인)

위 원 _____ 이영기 _____ (인)

초 록

Neural Machine Translation (신경망을 이용한 기계 번역) 모델을 학습시키는데 있어서 가장 영향을 많이 끼치는 요소는 학습 데이터인 병렬 말뭉치(Parallel Corpora)의 품질이다. 따라서 병렬 말뭉치의 품질 개선이 필수적이며 지금까지 다양한 정제(Refinement) 작업이 많이 도입되었으나 여전히 개선할 부분이 많다. 이 논문은 기계 번역 학습시 필요한 병렬 말뭉치의 품질 개선 작업에 도움이 될 수 있는 시각적 분석 시스템을 소개한다. 우리 시스템은 병렬 말뭉치의 Noise를 빠르게 발견하고 선별하기 위해 머신러닝 기술을 활용하여 다양한 지표(Metric)를 추출하고 이를 기반으로 상호작용이 가능한 시각적 분석 기법을 제공한다. 사용자는 우리의 시스템을 통해 Noise Data를 손쉽게 파악하고 이에 대한 상세한 내용을 확인 후 제거가 가능하다. 본 시스템의 효율성 및 유용함을 증명하기 위해 4명의 전문가를 포함한 총 8명의 사용자에게 사용성 평가를 진행하였으며, 마지막에 평가 결과를 바탕으로 개선해야 할 점에 대한 논의점도 언급한다.

주요어 : 시각적 분석 시스템, 신경망 기반 기계번역, 데이터 정제,
학 번 : 2021-29094

목 차

제 1 장 서 론 1

제 2 장 관련연구 4

제 3 장 디자인 요구사항 7

제 4 장 데이터 전처리 과정 10

제 5 장 시각화 디자인 14
- 제 1 절 Distribution View 14
- 제 2 절 Ranking View 15
- 제 3 절 Text Compare View 18
- 제 4 절 Ruleset View 20

제 6 장 사용성 평가 22
- 제 1 절 결과 23
- 제 2 절 사후 인터뷰 25

제 7 장 논 의 28

제 8 장 결 론 31

참고문헌 32

Abstract 36

감사의 글 37

그림 목차

[그림 1] 시각적 분석 시스템 개요 6
[그림 2] 데이터 전처리 과정 12
[그림 3] Distribution View <-> Ranking View 사용법 17
[그림 4] Text Compare View 사용법 19
[그림 5] Ruleset View 사용법 21
[그림 6] Universial Sentence Encoder Architecture 29

표 목차

[표 1-1] 데이터 전처리를 통해 추출한 9가지 지표 List 13
[표 2-1] 사후 인터뷰용 설문지 27

제 1 장 서 론

신경망을 이용한 기계 번역 학습(Neural Machine Translation,NMT)를 학습시키기 위해서는 2개의 언어 쌍으로 되어 있는 말뭉치(Corpus)로 구성된 Data를 필요로 한다. 흔히 이러한 쌍으로 구성된 말뭉치들을 Parallel Corpora 라고 하는데 병렬적으로 구성된 각 언어의 쌍(Pair)이다. 예를 들면 영어로 “I am a boy” 라는 문장이 있다면 한국어로 “나는 남자입니다” 라는 2개의 문장으로 구성된 쌍의 집합을 병렬 말뭉치(Parallel Corpora) 라고 정의하며 이는 NMT 학습에 사용된다.

일반적으로 NMT Model을 학습시키기 위한 병렬 말뭉치는 매우 큰 Data를 요구하기 때문에 Web이나 디지털화 되어 있는 책들을 통해 Crawling 기법을 활용하여 Data 추출을 진행한다. 다만 필연적으로 이렇게 추출된 데이터에는 Noise가 포함되어 있는 병렬 말뭉치를 포함할 수밖에 없다. 문제는 이러한 Noise 병렬 말뭉치가 잘못된 기계번역 학습을 유도한다는 것이다 [1].

결국 이렇게 잘못된 형태로 포함되어 있는 Noise 병렬 말뭉치는 NMT Model의 성능에 매우 큰 영향을 미치게 되기 때문에, 매우 낮은 품질 수준을 가지고 있는 병렬 말뭉치를 정제(Refinement)를 하는 것이 매우 중요한 역할을 한다. 따라서 NMT Model의 품질을 개선하기 위해서는 수많은 쌍으로 이루어진 병렬 말뭉치 (Parallel Corpora) 중에서 Noise인 것을 빠르게 찾아내고 이를 제거하는 것을 필요로 하는 것은 자명하다. 실제로 기계 번역 학습 작업(Task)에서 최고 학회로 여겨지는 EMNLP에서도 Noise 병렬 말뭉치를 Filtering 하기 위한 Shared Task가 별도로 존재할 정도로 최근에 이에 대한 작업의 중요성을 더욱 강조하고 있다 [2].

최근까지도 병렬 말뭉치를 정제하고 품질(Quality)를 향상시키기 위한 다양한 방법들이 소개되었다. 우선 사용자가 직접 모든 병렬 말뭉치를 하나씩 검사하면서 Noise를 찾는 방법이 있을 수 있다. 그러나 이렇게 사용자가 직접적으로 Noise를 검사하는 방법은 매우 시간이 오래 걸릴 뿐만 아니라, 엄청난 노동력을 필요로 하며, 수백만, 수천만 혹은 수억 개의 쌍으로 이루어져 있는 NMT 학습용 병렬 말뭉치를 전체를 검사하는 것은 현실적으로 불가능하다. 뿐만 아니라 이러한 직접 검사 방법은 사용자가 병렬 말뭉치를 구성하고 있는 2개의 언어 중 한 개만이라도 배경 지식이 없을 경우 Noise를 판별하는 것이 사실상 불가능하다. 이러한 단점을 극복하기 위해 소프트웨어 도구를 활용하여 자동으로 병렬 말뭉치의 Noise를 찾아내는 방법을 사용할 수 있는데, 대표적으로 규칙 기반(Rule-based)의 필터링(Filtering) 기술 활용하여 Noise를 제거하는 것이다. 필터링이 기준이 되는 규칙은 2개의 언어 쌍의 길이 차이 혹은 단어 개수 등의 차이를 등을 활용할 수 있으며 [3], 최근 인공지능 기술의 발전에 따라 AI 모델을 기반으로 문장의 의미적인 정보 (Semantic Information, 예: cosine similarity)를 추출하여 언어 쌍의 차이를 바탕으로 Noise를 찾는 방법도 소개되고 있다 [4]. 하지만 이러한 자동화된 도구를 바탕으로 병렬 말뭉치의 Noise를 추출하는 작업에 대한 단점도 존재하는데 대표적으로 '잘못된 분류(mis-classification)' 인 사례가 발생할 수 있다는 것이다 [5]. 예를 들면 언어 쌍에 의미론적인 문제는 없는데 단순히 미리 정한 규칙을 벗어난다는 이유로 Noise로 분류할 수 있다는 것이다. 그리고 언어의 쌍의 특성이 매우 상의한 경우(예: 영어 <-> 한국어) 규칙 기반의 필터링 방법은 위에서 언급한 '잘못된 분류'를 할 수 있는 확률이 높아지기에 다양한 타입의 Noise 병렬 말뭉치에 대한 선별 작업에 어려움을 겪을 수 있다 [1].

이러한 한계와 단점들을 극복하기 위해서, 우리는 신경망 기반 기계 번역 학습용 데이터인 병렬 말뭉치를 시각적 분석 시스템을 통해 정제하는 기법에 대해 소개한다. 우리의 시스템은 매우 큰 사이즈의 병렬 말뭉치의 전체적인 상태를 이해할 수 있도록 다양한 시각적 정보를 제공한다. 사용자는 우리가 제공하는 시스템을 통해 보다 효율적으로 Noise 병렬 말뭉치의 후보들을 여러가지 지표들 바탕으로 쉽게 선별 및 분류를 할 수 있으며, 1차적으로 선별된 Noise 후보군들 중 실제 Noise 병렬 말뭉치를 선별하고 이를 분석하여 추후 더 많은 Noise를 찾을 수 있도록 패턴(Pattern)을 파악할 수 있다.

본 논문에서는 우리가 만든 시각적 분석 시스템이 얼마나 효율적이고 유용한지에 대해 대기업에서 근무를 하고 있는 4명의 전문가를 포함하여 총 8명을 대상으로 사용성 평가를 통해 증명한다. 마지막으로 사용성 평가 결과를 분석하여 추후 논의해야 사항을 언급하며 마무리한다.

제 2 장 관련연구

먼저 관련연구에서는 도구를 활용하여 자동적으로 병렬 말뭉치에서 Noise를 선별하는 관련 연구를 소개한다.

초기에 병렬 말뭉치를 자동적으로 정제하는 작업은 일반적으로 쉽게 추출할 수 있는 문장의 속성 그 자체를 활용하였다. 예를 들면 Moses[3]의 연구에서는 원본 문장과 대상 문장 간의 길이 차이를 기반으로 문자 쌍을 필터링 하는 기법을 소개하였다. 다만 이러한 기존 접근 방식에서는 단순한 문장의 속성 차이만을 대상으로 하고 문장의 의미론적인 정보는 추출할 수 없기 때문에 모델을 기반으로 필터링 하는 방법이 추가로 제시되었다. 예를 들면, Xu et al. [7] 및 Zhang et al. [4] 연구의 경우 먼저 인공지능 기술을 활용하여 사전 훈련된 모델을 바탕으로 Source 및 Target 문장에 대해 임베딩 벡터(Embedding Vector)로 변환하고 해당 벡터간의 Cosine 유사도(Similarity)와 같은 두 문장 쌍의 실제 의미론적인 유사도 지표를 추출하여 병렬 말뭉치에서 Noise를 찾아내는 작업을 수행하였다. 그러나 이러한 자동 선별 도구를 활용한 접근 방식은 여전히 '잘못된 분류(Mis-Classification)' 문제로 어려움을 겪을 수 있다 [5]. 이러한 기존 연구의 한계를 극복하기 위해 우리가 제공하는 대화형 시각적 분석 시스템을 통해 병렬 말뭉치를 검사하고 개선하기 위한 방법을 필요로 하며, 우리의 시스템을 활용하여 사용자가 Noise 병렬 말뭉치의 후보군을 파악하고 실제 Noise를 제거하는 과정이 필요하다는 것에 대한 요구사항이 존재한다는 것을 확인할 수 있다.

다음으로 신경망을 이용한 기계 번역 학습과 관련된 시각적 분석

시스템의 선행 연구에 대해 살펴본다. 우선 Data2Vis[8]는 Sequence to Sequence 언어 모델을 이해하기 위한 대화형 시각화 시스템을 제공하였다. 다만 해당 연구는 기계 번역 학습을 위한 Dataset을 시각화한 것이 아닌 단순한 모델 자체를 분석하기 위한 시스템이기 때문에 본 논문에서 진행하려는 연구와는 큰 상관관계가 없다. 일부 연구의 경우 번역 프로세스에 대한 자세한 설명을 위해 Attention 점수를 시각화하는데 중점을 둔 연구도 존재한다 [9]. Munz et al.[10]는 신경망 기계학습 모델에서 제공하는 단일 언어 데이터의 번역 결과를 검토하고 사용자가 잘못된 번역을 수정할 수 있도록 하는 시각적 분석 시스템을 제안하였다.

그러나 지금까지 신경망 기반의 기계 번역 학습(NMT) 모델에 대한 시각적 분석 시스템이 많이 제안되었지만, 모델 자체를 대상으로 삼았으며 학습용 Dataset인 병렬 말뭉치에 대한 시스템은 아직 도입된 사례를 찾지 못하였다. 기존 연구 중에 mixed-initiative 인터페이스 방식의 대화형 데이터 랭글링(Data Wrangling) 도구 [11]를 활용할 수 있지만, 이는 병렬 말뭉치를 처리하는데 있어 적합하지 않다는 문제가 있다. 따라서 우리는 병렬 말뭉치를 효과적으로 검사하고 정제하는데 있어 특화된 대화형 시각적 분석 시스템을 도입하여 기존 연구의 한계점들을 극복하고자 하는 것을 목표로 한다.

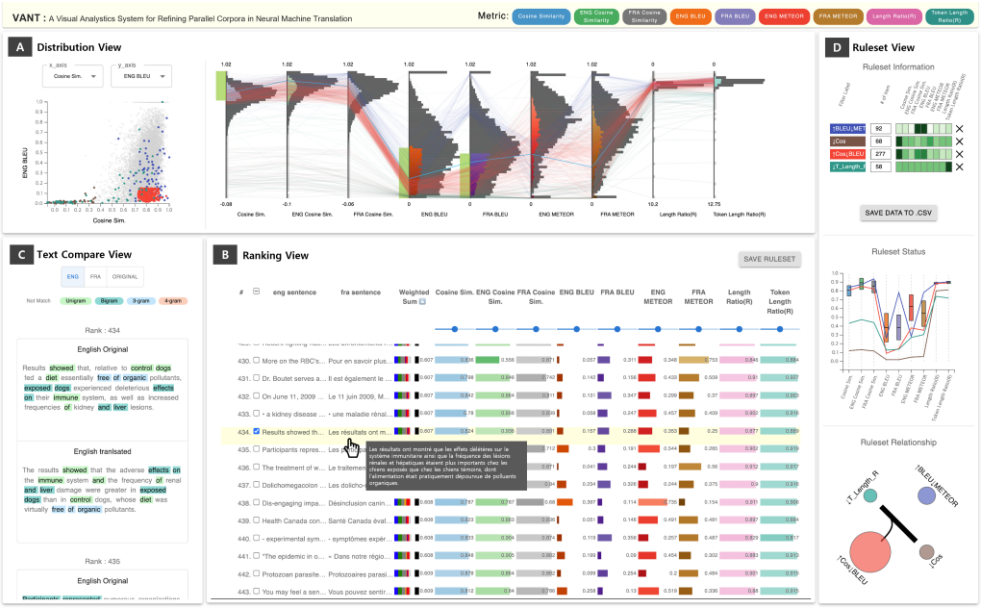

그림1. 기계번역학습 데이터셋 품질 평가를 위한 시각적 분석시스템의 개요. (A) Distribution View는 각기 다른 종류의 지표들을 통해 병렬 말뭉치의 품질에 대한 개요를 제공한다. (B) Ranking View에서는 사용자가 각 지표의 가중치를 조절하여 관심있어 하는 Noise 병렬 말뭉치의 후보군들 정렬하여 나타낸다. (C) Text Compare View는 사용자가 실제 후보군으로 판단한 Noise 병렬 말뭉치들에 대한 상세 내역을 확인할 수 있고 실제 얼마나 문장 쌍이 일치하는지를 시각적으로 표현해준다. (D) Ruleset View에서는 Noise로 선별한 병렬 말뭉치들이 어떠한 통계적인 특성을 가지고 있는지를 시각적으로 표현하고 이를 통해 패턴을 파악한 후 더 많은 Noise를 찾을 수 있도록 한다.

제 3 장 디자인 요구사항

우리는 신경망을 이용한 기계 번역(NMT)의 학습 데이터로 사용되고 있는 병렬 말뭉치의 정제 작업이 실제 작업에서 얼마나 큰 영향과 중요도를 가지는지에 대해 판단하기 위해 8년 이상의 경력을 가진 2명의 전문 엔지니어와 예비 인터뷰를 진행하였다. 인터뷰에서는 우리는 주로 (1) Noise가 있는 문장 쌍을 찾고 병렬 말뭉치를 정제해야 하는 필요성이 얼마나 있는지 (2) 데이터 필터링의 과정에서 겪는 어려움이 어떠한 것들이 있는지에 대해 중점적으로 논의를 진행하였다. 우리는 본 인터뷰를 통해 실제 말뭉치 정제 작업이 현재 어떤 식으로 이루어져 있으며 어떤 어려움을 겪고 있는지 파악할 수 있었으며 인터뷰 내용을 바탕으로 우리가 개발하려는 시각적 분석 도구에 대한 4가지 디자인 요구사항(Design Requirement, DR)을 도출할 수 있었다.

DR1: NMT 데이터 품질을 개선하기 위한 전반적인 개요를 제공

NMT 모델을 훈련시키기 위해서는 매우 큰 규모(Size)의 병렬 말뭉치를 필요로 하지만 이 중에서 Noise 병렬 말뭉치를 식별하는 것은 매우 많은 노력을 필요로 한다. 따라서 우리의 시스템은 사용자가 병렬 말뭉치를 쉽게 이해하고 잠재적으로 실제 Noise가 될 수 있는 문장 쌍들의 후보군을 쉽게 찾을 수 있도록 문장 자체를 표현하는 것이 아닌 지표(Metric) 등을 활용하여 품질 점수의 분포를 표시해야 한다.

DR2: 사용자가 다양한 지표의 조합하여 Noise 후보군을 추천

각각의 지표는 단일 관점에서만 병렬 말뭉치의 품질을 나타낼 수밖에

없다. 따라서 사용자는 보다 포괄적으로 Noise 후보를 찾기 위해 여러가지 지표들을 조합하여 전체적인 병렬 말뭉치 품질을 나타낼 수 있도록 해야 한다. 따라서 우리의 시스템은 여러 지표를 제공하고 사용자가 Noise 후보군을 가능성이 높은 순으로 정렬하기 할 수 있도록 하기 위해 지표의 가중치를 대화식 인터페이스를 통해 조정하도록 해야 한다. 이를 통해 사용자는 다양한 관점에서 Noise 후보군을 조사할 수 있으며 여러 조합을 통해 실제 품질이 낮은 병렬 말뭉치를 쉽게 찾을 수 있다.

DR3 : 언어 배경지식 없이도 Noise 병렬 말뭉치를 판단 가능

사용자는 병렬 말뭉치를 구성하고 있는 두 언어에 모두 능통하지 않다면 자연어 형태의 원본(Source)와 대상(Target) 문장의 Noise를 판단하는 것이 매우 어렵다. 따라서 우리의 시스템은 사용자가 언어적인 배경이 없어도 Noise를 판단할 수 있도록 시스템에서 자동적으로 사용자가 선호하는 언어로 두 문장의 비교가 가능하도록 하는 기능이 필요하다.

DR4: 분류된 Noise를 바탕으로 Noise 말뭉치의 패턴 분석 지원

Noise가 많은 병렬 말뭉치는 다양한 형태로 존재할 수 있다. 따라서 사용자는 더 많은 Noise를 찾기 위한 후보군을 찾기 위해 이전에 식별(Identify)된 항목의 정보를 바탕으로 정보를 분석을 하는 것이 중요한 작업이다. 본 시스템은 사용자가 이전에 선택했던 이력(History)를 추적할 수 있을 뿐만 아니라 식별된 Noise의 특성을 분석하여 패턴을 발견할 수 있도록 지표 점수 및 가중치 등에 대한 실제 Noise 통계 정보를 제공해야 한다. 사용자는 추후 패턴을 파악한 후 이를

적용하여 추가적인 Noise가 병렬 말뭉치의 찾을 수 있다.

제 4 장 데이터 전처리 과정

우리는 디자인 요구사항1(DR1)을 만족시키기 위해 병렬 말뭉치의 품질을 반영하는 여러가지 지표(Metric)를 준비하였다. 지표의 가중치는 대화식 인터페이스를 통해 조정이 가능하고 병렬 말뭉치 중 Noise의 가능성이 높은 순위들을 1차적으로 확인할 수 있다. 사용자는 다양한 지표를 통해 여러 관점에서 말뭉치들을 효과적으로 검사 (Inspection) 할 수 있다. 다만 이러한 사용자의 작업을 가능케 하려면 다양한 지표를 추출하는 것이 필수적으로 필요한데 이를 위한 전처리 과정은 아래와 같다.

우리는 먼저 자연어 형태의 문장에서 일반적으로 쉽게 추출할 수 있는 길이(Length) 및 토큰 길이 비율(Token Length Ratio)를 추출하였다. 다만 이러한 일반적인 속성들은 문장의 의미론적인 관점에서의 유사성을 나타낼 수 없으며, 언어의 태생이 다른 경우(예: 한국어 및 영어) 유용한 정보가 되지 않을 수 있기 때문에 한계를 지니고 있다.

이러한 단점을 보완하기 위해 우리는 언어의 유형 관계없이 의미론적인 문장의 정보를 추출하기 위해 임베딩 벡터(Embedding Vector)로 인코딩하는 방법을 채택하였다. 이를 위하여 인공지능 기술을 바탕으로 사전에 훈련된 모델 인코더인 Universal Sentence Encoder[12]를 활용하여 문장을 임베딩 벡터로 변환하였다. Universal Sentence Encoder 모델은 언어에 상관없이 다양한 언어를 임베딩 벡터로 변환이 가능하기에 임베딩 벡터로 변환된 2개의 벡터를 바탕으로 2개의 문장 간의 Cosine 유사도(Similarity)를 추출할 수

있었다.

뿐만 아니라 Target 문장에서 Source 문장으로 변환하여 데이터를 증가시키고 이를 신경망을 이용한 기계 번역 학습에 이용하는 역번역 기법(Back-Translation)[13]에 영감을 받아 우리는 Google Translation API를 사용하여 Source 와 Target 언어를 Target과 Source 언어로 번역하였다. 위와 같은 작업을 통해 우리는 2가지의 병렬 말뭉치를 평가할 수 있는 지표(Metric)인 BLEU[14] 및 METEOR[15]를 추출할 수 있었다. 2가지 지표는 n-gram[16]의 정밀도(Precision)와 재현률(Recall)의 조화(Combination)을 바탕으로 추출된 지표로 실제 두 문장의 쌍의 일부가 얼마나 일치하는지 비교하는데 있어 자연어처리 학계 영역에서 범용적으로 사용되는 방법이다. 역번역 결과는 추후 본 논문에서 언급된 텍스트 비교 보기에서 제공되므로 Source 또는 Target 언어에 대한 전문 지식이 부족하거나 배경 지식이 없는 사용자도 우리의 시스템을 통해 Noise 여부를 판별할 수 있다(디자인 요구사항3, DR3).

최종적으로 우리 시스템에서 위에서 언급된 기술들을 바탕으로 총 9가지 지표를 추출하였으며 목록은 붙임과 같다 [표1]. 또한 데이터 전처리를 위한 파이프라인은 그림2에 간략화 하여 묘사하였다.

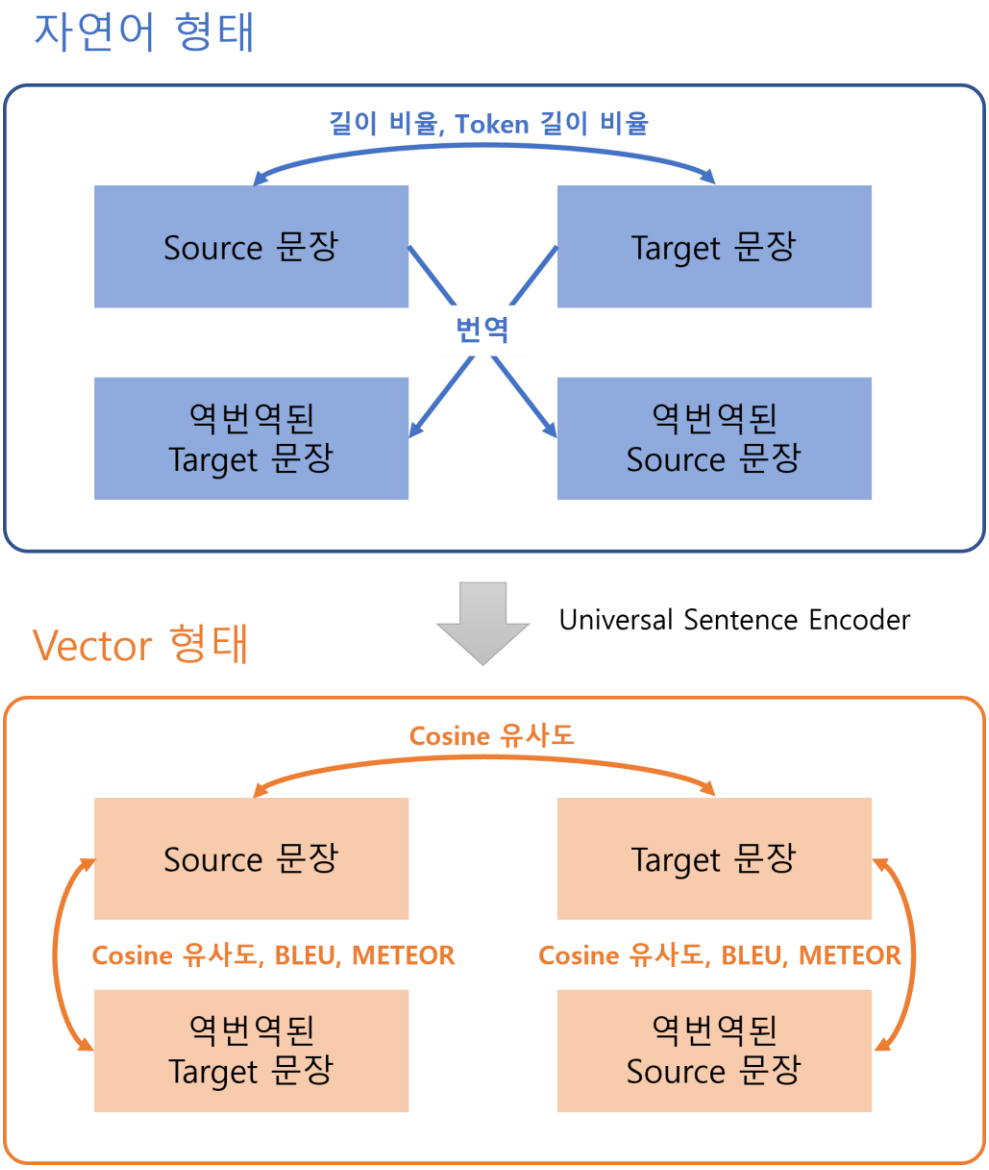

그림2. 본 시스템의 데이터 전처리 과정을 묘사한 그림. 병렬 말뭉치의 품질을 나타낼 수 있는 지표를 추출하기 위하여 자연어 형태를 Universal Sentence Encoder를 이용하여 Embedding Vector로 변환 후 여러 지표들을 추출한다.

지표 이름	Description
Cosine similarity	Text Similarity between original source and target corpus
Length Diff Ratio	Length Ratio based on source corpus
Token Length Diff Ratio	Tokenized word Length Ratio based on source
Back-translated Cosine Similarity (Source to Target)	Cosine similarity between source and back-translated target language to source
Back-translated Cosine Similarity (Target to Source)	Cosine similarity between source and back-translated source language to target
Source BLEU	BLEU score between source and back-translated source
Target BLEU	BLEU score between target and back-translated target
Source METEOR	METEOR score between source and back-translated source
Target METEOR	METEOR score between target and back-translated target

표1. 데이터 전처리 과정을 통해 추출한 9가지의 지표

제 5 장 시각화 디자인

우리는 제3장에서 추출한 디자인 요구사항을 충족시키기 위해 대화형 시각화 시스템을 개발하였다. 그림1과 같이 우리의 시스템은 Distribution View, Ranking View, Text Compare View 및 Ruleset View 총 4개의 보기(View)로 구성하였다.

각 View에 대한 자세한 설명에 앞서 사용자 관점에서 본 시스템을 사용하는 일반적인 순서는 아래와 같이 진행된다. 먼저 Distribution View를 통해 지표 점수를 기반으로 병렬 말뭉치 내에서 Noise 후보군을 1차적으로 선택한다. 두번째로 Ranking View와 Text Compare View를 이용하여 선발된 Noise 후보자들의 상세 내역을 확인한다. 세번째로 실제 Noise 말뭉치 쌍을 Ranking View를 통해 확인하고 실제 Noise로 판별된 문장 쌍을 선택하면 이를 Ruleset의 집합으로 저장한다. 마지막으로 Ruleset View에서 실제 Noise가 있는 문장 쌍의 패턴을 분석한다.

5.1 Distribution View

Distribution View(그림 1A)에서 사용자는 NMT 학습용 병렬 말뭉치의 특성을 빠르게 식별하고 Noise가 많은 병렬 말뭉치에 대한 단서를 얻을 수 있다(디자인 요구사항1, DR1). 우리는 이를 위해 2가지의 시각적인 구성 요소를 제공한다. 1) PCP(Parallel Coordinate Plot)는 전처리 단계에서 추출된 각 지표 점수 간의 관계를 경로(Path) 형태로 표현하는 시각적 기법을 사용한다. 2) 산점도(Scatter Plot)를 바탕으로 x, y 축을 사용자가 선택한 2가지 지표에 대해 데이터 분포를 표현한다. 그러나 우리가 다루고 있는 병렬 말뭉치는 크기가 일반적으로

매우 크기 때문에 PCP에서 모든 지표들 간의 경로(Path)를 표현하는 것은 사실상 불가능하며 표현을 한다 하더라도 사용자에게 시각적인 혼란을 야기시킬 수 있다. 이 문제를 해결하기 위해 우리는 Parallel Histogram Plot(PHP) [17] 시각적 인코딩 방식을 채택하여 각 축이 각 지표의 분포를 각기 다른 색상으로 나타낼 수 있도록 히스토그램 기법을 결합하여 사용자의 시인성에 대한 개선을 진행하였다.

PHP와 함께 제공되는 산점도(Scatter Plot)을 통해 2개의 지표 간의 상관 관계를 더 상세히 확인할 수 있으며, 사용자는 PHP내의 각 지표를 나타내는 하나의 축에서 브러싱(Brushing) 기법을 활용하여 Noise 후보를 필터링 할 수 있도록 지표 값들의 범위를 대화형 인터페이스를 통해 선택할 수 있다. 그림3과 같이 선정된 후보를 Ranking View에서 표현함으로써 사용자가 실제 지표에 대한 세부사항을 확인하는데 도움을 준다(디자인 요구사항1, DR1). 추가적으로, 사용자는 PHP의 각각의 축을 Drag & Drop 하여 축 순서를 변경할 수 있다. 축 순서의 변경은 Ranking View 및 Ruleset View와 같은 다른 모든 View들의 순서와 연동된다.

5.2 Ranking View

Ranking View (그림 1B)는 Distribution View에서 선택된 Noise 후보 순위에 대해 자세한 지표(Metric) 정보를 사용자가 열람할 수 있도록 시각적인 요소 제공한다. 앞에서 지속적으로 언급한 것처럼 우리가 다루고 있는 병렬 말뭉치의 크기는 사용자가 하나씩 하나씩 탐색하기에는 사이즈가 매우 크기 때문에 사용자에 의한 직접적인 탐색은 사실상 불가능하다. 따라서 우리는 사용자가 Noise를 보다 빠르고 쉽게 감지할 수 있도록 다양한 지표들을 가중치에 따른 점수의 총 합계를 바탕으로 우선순위를 추출하고 이를 표현한다 [18].

Lineup[19]의 시각화 도구로부터 영감을 받아, 우리는 각 지표의 가중치를 조정하기 위한 Slider Bar를 제공하며 이에 대한 자세한 정보를 보여주는 Table(표)를 보여준다. 우선 가중치가 설정되면 Ranking View는 각각의 개별 지표 점수를 기반으로 각 후보들 간의 가중치 합계를 계산한 후 이를 오름차순 혹은 내림차순으로 정렬하여 보여준다(디자인 요구사항4, DR4).

Ranking View에서 제공하는 표의 각 행은 (1) 병렬 말뭉치를 구성하고 있는 실제 자연어 형태의 Text, (2) 가중치의 합계, (3) 개별 지표 점수를 보여준다. 가중치의 합계는 Stacked Bar 형태로 표시되며 각 지표 점수는 각기 다른 색상을 가진 하나의 막대로 표현된다. 막대의 길이는 지표의 점수를 나타내고, 채도(Saturation)는 순위를 나타낸다.

사용자가 표에서 마우스를 가져가면 Text Compare View(그림 1c)에서 세부 사항 확인을 위해 마우스가 올려진(Hovering) 항목에 해당하는 말뭉치 쌍으로 이동하게 되고 PCP에서는 해당 항목과 관련된 경로에 대해 Highlighting 기법을 통해 Path의 색상을 더 진하게 표현한다. 사용자는 Ranking View에서 정렬된 항목을 하나씩 살펴보고 검사함으로써 각 후보자가 실제 Noise인지 아닌 지에 대해 판별을 할 수 있게 된다. 최종적으로 Noise로 판별된 항목에 대해 Check Box를 통해 항목을 선택한 후 저장 버튼(Save)를 클릭하며 Noise 후보자들의 Ruleset을 저장할 수 있다. 사용자는 Ruleset을 저장할 때 색상과 이름을 지정할 수 할 수 있으며 시스템은 해당 정보를 통해 Ruleset을 저장한다.

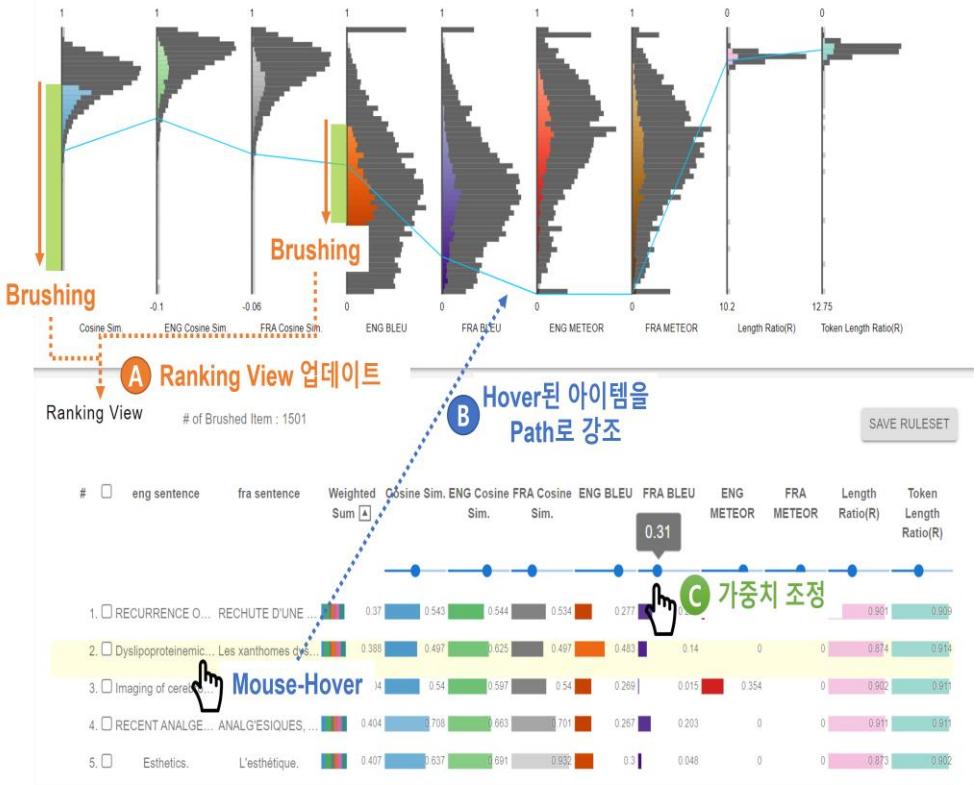

그림3. Distribution View와 Ranking View의 연결 관계를 표현한 그림. (A) 사용자는 PCP 상의 여러 개의 축을 Brushing하여 범위를 선택할 수 있으며 해당 범위의 상세 내용은 Ranking View에서 표시된다. (B) Ranking View에서 아이템을 Mouse-Hover할 경우 PCP에 해당 아이템이 강조되어 표시된다. (C) 가중치를 조정하게 되면 Ranking View의 항목이 자동적으로 업데이트 된다.

5.3 Text Compare View (텍스트 비교 View)

우리의 시스템은 병렬 말뭉치를 평가하기 위한 다양한 지표가 제공되지만 이는 어디까지나 품질에 대한 지표 자체일 뿐 실제 말뭉치의 품질을 완전하게 반영하는 것은 아니다. 따라서 사용자는 병렬 말뭉치의 원본데이터(Raw Data)를 확인하고 검사해야 한다. 우리의 디자인 요구사항은 Source 혹은 Target 언어 중 하나라도 언어적인 배경지식이 없는 사용자도 고려하기 때문에(디자인 요구사항3, DR3) Text Compare View에서 (그림 1c)는 3가지 타입에 대한 언어 선택 옵션 Source <-> Target, Target<-> Source를 제공한다. 사용자는 Source 또는 Target으로 표현되어 있는 Text Box를 클릭하게 되면 Text Compare View에서는 선택된 언어로 Source 혹은 Target 언어가 역번역(Back-Translated)된 항목으로 변환하여 표시한다. 예를 들어 병렬 말뭉치가 영어와 프랑스어로 구성되어 있는 경우 사용자가 '영어' 버튼을 클릭했을 경우 프랑스어는 영어로 자동적으로 번역되어 사용자는 Source와 Target 언어가 모두 영어로 표현된 데이터를 참조할 수 있게 된다. 마지막으로 사용자가 표현하고 싶은 언어를 클릭할 경우 말뭉치가 정제되지 않은 Text 형태로 표현되면서 Source와 Target 언어에서 N-Gram 매칭이 얼마나 되었는지에 대해 배경색을 통해 시각적으로 판단할 수 있도록 해준다. N-Gram 매칭은 실제 2개의 문장의 유사성을 나타내는 방식으로 2개의 문장의 공통점과 차이점을 보다 쉽게 파악할 수 있으며, Bigram, 3-gram, 4-gram은 각기 다른 Text 배경색으로 강조되어 표현된다.

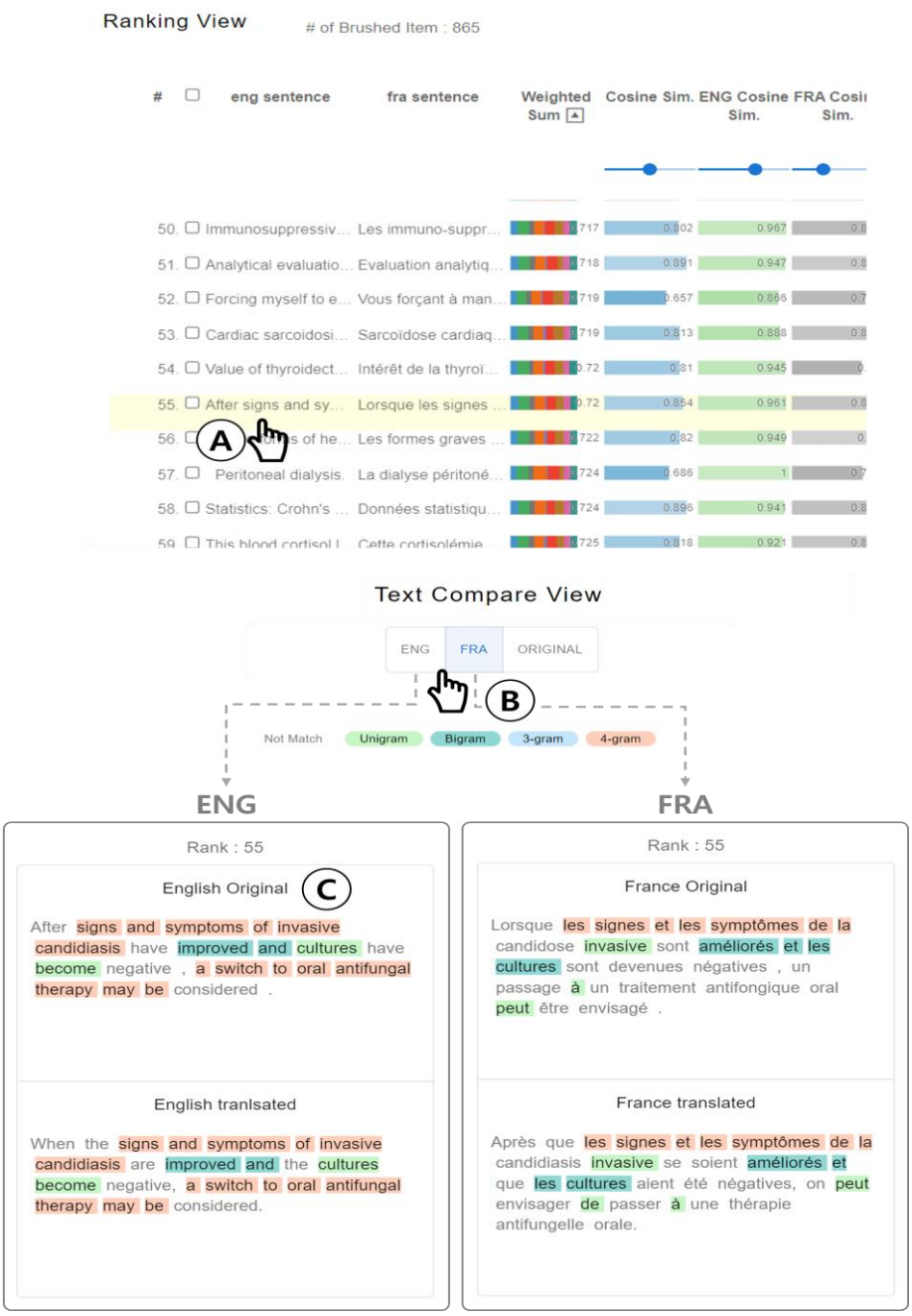

그림4. (A) 사용자가 Ranking View에서 아이템에 Hover를 하게 되면 Text Compare View는 자동적으로 해당 문장 쌍을 표현한다. (B) 사용자는 선호하는 언어를 선택할 수 있다. (C) 문장의 배경색은 2개의 문장이 얼마나 n-gram 매칭이 되어 있는지를 표시한다.

5.4 Ruleset View

일단 사용자가 Noise 병렬 말뭉치에 후보군에 대해 Ranking View와 Text-Compare View를 통해 선별하면 Ranking View에서 제공하는 Save 버튼을 클릭 후 이를 Ruleset 형태로 저장할 수 있게 된다. 이후 사용자는 더 많은 Noise를 찾기 위해서 저장된 Ruleset에 대한 통계 정보를 바탕으로 패턴을 파악하고 추후 이를 활용할 수 있다. 일단 Ruleset이 만들어 지면 선별된 Noise 문장 쌍들이 사용자가 선택하여 저장된 색상으로 Distribution View에 항시 강조되어 표시되게 된다. 사용자에게 패턴 분석을 더 쉽고 용이하게 할 수 있도록 Ruleset View는 3개의 하위 View로 구성되어 제공된다. 1) Ruleset Information View 2) Ruleset Status View 3) Ruleset Relation view.

먼저 Ruleset Information View에서는 각각의 Ruleset의 Meta Data 정보를 표현한다. Meta Data는 사용자가 지정한 이름, 색상, Cardinality 및 각 지표의 가중치로 구성되어 있으며 이를 각각의 행으로 표현한다 (그림 5A). 각 지표의 가중치는 Ruleset 간의 직관적인 비교를 위해 Heatmap 형태로 표현된다. 사용자가 우선 행을 클릭하면 Ranking View 및 Text Compare View에서 Ruleset으로 저장된 병렬 말뭉치들이 자동적으로 선택되어 이를 나타나게 된다. 또한 Distribution View 내부에 존재하는 PHP 및 산점도(Scatter Plot)에서도 해당 항목들이 강조되어 동시에 표시되게 된다. 이를 통해 사용자는 이미 저장한 각각의 Ruleset에 대한 기록(History)를 쉽게 추적할 수 있다. Ruleset Status View에서는 Ruleset에서 저장된 각 지표의 점수의 평균을 Line Graph로 표현하게 되며 해당 점수의 최대값과 최소값의 통계를 나타낼 수 있도록 Box Plot을 제공한다. 마지막으로 Ruleset Relation View에서는 각각의 Ruleset간 선택된 Noise 병렬 말뭉치가

중복되어 있는지에 대해 시각적으로 표현하게 된다(그림 5B). 해당 View에서는 하나의 원이 Ruleset에 저장된 병렬 말뭉치의 수를 크기에 따라 표현하게 되고, 2개의 Ruleset이 공통 항목이 있을 경우 선을 표시하여 연결시키며 선의 너비(Width)는 공통으로 선택된 병렬 말뭉치 수를 나타낸다.

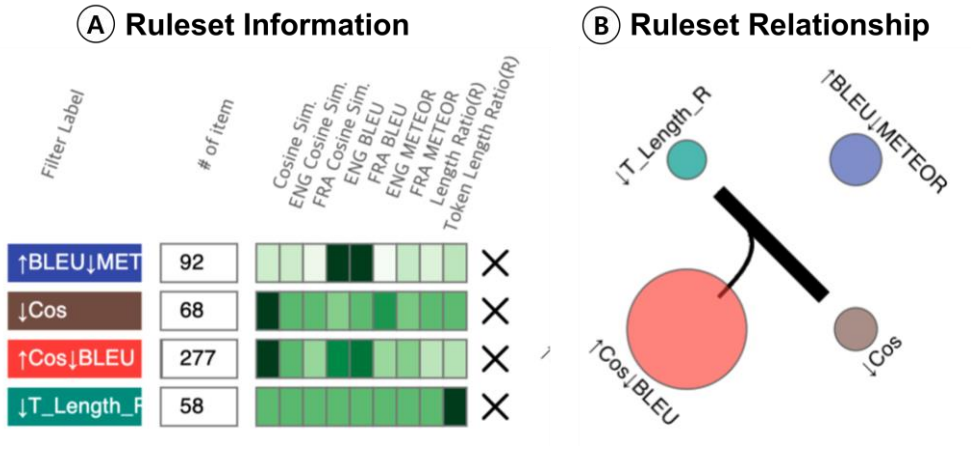

그림5. (A) Noisy로 선별된 병렬 말뭉치들은 Ruleset으로 저장되며 사용자가 지정한 이름 및 색상에 따라 Label을 저장한다. 이때 사용자가 지정한 가중치는 Heatmap을 통해 표현된다. (B) Edge Bundling을 통해 각 Ruleset 같은 관계를 표현한다. 이 경우 ↑Cos↓BLEU, ↓Cos, ↓T Length R에서 공통적으로 선택한 아이템이 있다는 것을 알 수 있다.

제 6 장 사용성 평가

우리는 디자인 요구사항에 맞춰 개발된 시각적 분석 시스템의 효과와 유용함을 증명하기 위해 Samsung Research에서 근무하고 있는 8명의 연구원을 대상으로 사용자 연구(User Study)를 진행하였다. 참여자는 실제 Natural Language Processing 팀에서 근무하고 있는 4명의 전문가 (E1- E4)와 Software Engineering 팀에서 근무하고 있는 4명의 전문 소프트웨어 엔지니어 (S1-S4)로 구성되었다. 모든 참가자들은 Software 관련 업무를 6년 이상 했던 경험들을 가지고 있으며 한국어를 자유롭게 구사할 수 있으며 영어를 가급적 유창하게 할 수 있었으나, 프랑스어에 대한 배경 지식은 전혀 없는 상태였다.

우리 시스템을 평가하기 위해 실제 기계번역학습에서 빈번하게 사용되고 있는 실제 병렬 말뭉치 데이터 세트를 사용하였다.

1) Scielo Corpus [20] – 영어, 프랑스어로 구성되어 있는 의학 / 생물학 관련 병렬 말뭉치
2) 한국어 <-> 영어로 구성되어 있는 영어 뉴스 데이터 [21].

우리는 실제로 사용자가 언어적인 지식 배경이 없이도 시스템을 사용하는 방법을 관찰하기 위해 영어 <-> 프랑스어로 구성된 Scielo Corpus를 준비하여 우리 시스템의 효용성에 대해 검증하고자 하였다. 사용성 평가는 아래와 같은 단계를 통해 수행되었다.

1) 시스템의 목적 및 각 View에 대한 설계 의도 등에 대해 15분 동안 간략하게 사용자에게 설명.
2) 이후 20분 동안 실제 Noise 병렬 말뭉치를 어떻게 발견하고 선별하는 작업에 대해 재현하는 과정을 사용자에게 보여줌.

3) 사용자가 직접 20분 동안 Noise 병렬 말뭉치를 선별하는 작업에 대해 요청.

4) 실험이 끝난 후 우리의 시스템을 질적으로 평가하기 위한 사후 인터뷰를 진행.

6.1 결과

Noise 후보 발견 과정

모든 참가자들은 주로 Noise 후보군을 선택하기 위해(디자인 요구사항1) 화면 중앙에 표현되어 있는 Distribution View를 활용하여 각 지표의 점수 분포를 조사하였다. 이후 참가자들은 Parallel Coordinate Plot(PCP)내에서 브러싱(Brushing) 기법을 활용하여 자신이 관심이 있어 하는 지표에 대한 점수 범위를 선택하였다. 예를 들면 BLEU와 METEOR 점수가 낮은 항목을 선택하기 위하여 BLEU와 METEOR가 표현되어 있는 축에서 범위를 선택하였다. 상대적으로 지표에 대한 이해도가 높은 도메인 전문가의 경우 Distribution View 및 Ranking View에서 다양한 인터랙션을 통해 Noise 병렬 말뭉치 후보군을 찾기 위해 노력하는 모습을 볼 수 있었다. 예를 들면 전문가 E1, E2 및 E4는 PCP 내에서 여러가지 지표들을 반복적으로 선택하고 Ranking View에서 이에 대한 병렬 말뭉치가 실제 어떤 지표 점수를 가지고 구성되어 있는지에 대해 파악하고 가중치를 조정하였다. 대부분의 전문가들은 기계번역학습에서 대표적으로 사용되는 BLEU와 METEOR 지표의 가중치를 증가시키고 일반적인 속성으로 분류되는 문장 길이 및 Token 길이 비율(Token Length Ratio)의 가중치를 감소시키는 것을 확인할 수 있었다.

또한 일부 참가자(E3, S2, S3)의 경우 낮은 점수를 가지고 있는 Cosine 유사도 지표에 관심을 가지고 선택하였으나 실제 Text

Compare View를 통해 Data를 확인했을 경우 실제 Noise가 아닌 사례를 발견하기도 하였다. 이를 통해 낮은 점수를 가지는 Cosine 유사도가 반드시 낮은 품질의 병렬 말뭉치를 보장하지 않는 다는 사실을 발견할 수 있었다.

후보군으로부터 실제 Noise 병렬 말뭉치 검사

사용자가 Noise 병렬 말뭉치 후보군으로 선택한 항목에 대해서 실제로 Noise 병렬 말뭉치인지 확인하는 과정은 Text Compare View를 통해서 이루어졌다 (디자인 요구사항3, DR3). 대부분의 참가자들은 N-Gram 매칭을 기반으로 한 문장 쌍을 강조하여 배경색으로 표현된 시각적 인코딩 방법이 실제 Noise 병렬 말뭉치인지 여부를 판단하는데 매우 도움이 된다고 언급하였다. 특히 영어와 프랑스로 구성되어 있는 데이터셋의 경우 참가자들이 프랑스어에 대한 배경지식이 전혀 없음에도 불구하고 프랑스어가 영어 문장으로 역번역(Back-Translated)된 기법을 활용하여 영어 <-> 영어 형태로 비교할 수 있었기에 병렬 말뭉치를 더 쉽고 빠르게 비교할 수 있었다고 응답하였다.

Ruleset 저장 및 Noise 패턴 검사

사용자들은 Noise 병렬 말뭉치의 후보군 들로부터 실제 Noise를 선별하고 이를 분석하기 위해 Ruleset을 저장하고 패턴을 분석하였다. 참가자들의 절반 이상 (E1-E4 및 S1)은 Distribution View내 PCP에서 강조된 Noise 병렬 말뭉치의 시각적 표현이 이전 Ruleset 저장 기록을 추적하는데 유용하다고 언급하였다. 그 외에도 전문가인 E3 및 E4는 Ruleset으로 저장되어 있는 Noise 병렬 말뭉치가 Distribution View내 산점도(Scatter Plot)에서 어떤 식으로 분포가 되었는지 관심을 가지고

살펴보았다. E3의 경우 Ruleset Relationship View에서 자신이 어떤 말뭉치를 Noise로 분류하여 공통적으로 선택하였는지 연결되어 있는 선(Line)들을 선택하여 확인할 수 있었다. E4는 산점도 내에서 X, Y축을 변경하기 위해 지표들을 바꾸면서 실제 Noise가 어떤 패턴을 가지고 있는지에 대해 중점적으로 살펴보기도 하였다.

6.2 사후 인터뷰

우리는 각 사용자의 개별 인터뷰가 끝난 이후 시스템의 유용성과 개선사항을 파악하기 위해 별도의 사후 인터뷰를 진행하였다. 모든 참가자들은 Distribution View에서 표현되는 각 지표의 점수 분포도가 병렬 말뭉치 데이터의 품질을 전반적으로 이해하는데 도움이 된다고 언급하였다. 참가자들은 다양한 View와 시스템에서 제공하는 대화식 인터페이스를 통해서 Noise 병렬 말뭉치를 탐색하는 것이 수월하다고 말하였다. 뿐만 아니라 Ranking View에서 가중치를 조정하고 Noise로 선별된 말뭉치들의 Ruleset 정보를 시각적으로 표현하는 것이 Noise 패턴을 찾고 식별하는데 도움이 된다고 하였다.

대부분 참가자들(E1, E2 및 S1-S4)들은 Text Compare View에서 역번역(Back-Translation)으로 번역된 문장들이 Source 혹은 Target 언어에 대한 배경지식이 없더라도 Noise 여부를 판단할 수 있도록 해줬기에 병렬 말뭉치 검사에 소요되는 시간 비용을 크게 줄일 수 있을 것으로 기대한다고 하였다.

특히 이번 인터뷰 대상자들은 모두 Software 및 자연어처리 업무를 오랫동안 해온 엔지니어들로서 만약 우리가 개발한 시각적 분석 시스템을 현업에서 적용한다면 큰 도움이 될 것이라고 언급하였다. 일례로 E1, E2, E4 참가자의 경우 우리 시스템을 실제로 현업에서 사용

가능한지에 대해 문의하였으며 실제 배포 일정 등에 대해 상세히 질문하였다.

이 외에도 인터뷰 참가자들은 우리 시스템의 개선점에 위한 여러가지 제안을 하였다. 예를 들면 E3, E4는 지금 고정적으로 제공되고 있는 9개의 지표를 자유롭게 수정/삭제/추가할 수 있는 인터페이스를 제공하는 것의 필요성에 대해 언급하였다. 뿐만 아니라 E3, E4와 같은 전문가들은 일반 속성으로 구성되어 있는 지표보다는 자연어처리 분야에서 많이 사용되고 있는 ROGUE, Perplexity와 같은 지표를 더 추가한다면 Noise 병렬 말뭉치를 찾는데 더 큰 도움이 될 것이라 언급하였다. 전문가들과는 반대로 일반 Software 엔지니어(S1, S2, S3)는 Ranking View에서 가중치를 조정하는 부분에 대한 대화식 인터페이스가 사용하기 다소 어렵다고 하였으며 Ruleset View 내에서 Relationship View가 어떤 의미를 갖는지 다소 직관적으로 이해하기 어렵다고 말했다. 마지막으로 대부분 참가자들은 Text Compare View의 사용성 개선 사항에 대해 언급하였다. 그들은 해당 View가 실제 Noise 병렬 말뭉치를 판단하는데 있어 가장 중요한 역할을 하지만 상대적으로 작은 크기로 인해 사용하기가 어렵다고 하였다. 특히 Text Compare View는 가장 자주 사용되기 때문에 더 나은 시인성을 확보하기 위해 윈도우 크기를 동적(Dynamic)으로 늘릴 수 있는 “확장”(Expansion) 기능을 추가할 것을 제안하였다.

Machine Translation Visualization Tool 설문지

이름 :

1. 실제 Machine Translation에서 Data Noise Filtering은 꼭 필요한 과정인가요? (1~10점)

2. 실제 우리 Tool을 이용해서 Manual Noise Detection을 쉽게 할 수 있었나요? (1~10점)

3. 본 Tool에서 가장 활용도가 좋았던 / 나빴던 View가 각각 어떤 것인가요?
   A. Overview (Parallel Coordinate Plot)
   B. Ranking View
   C. Text Compare View
   D. Filter View

4. 해당 View가 좋았던 이유는?

5. 해당 View가 나빴던 이유는?

6. 본 Tool에서는 Back Translation에 아이디어를 얻어서 Language Context가 부족하더라도 Text를 직접 비교할 수 있도록 하였습니다. 얼마나 도움이 되었나요? (1~10점)

7. 의견을 자유롭게 남겨주세요

표2. 사후 인터뷰용 설문지

제 7 장 논 의

사용성 측면

인터뷰 결과 자연어처리에 대한 경험이 많고 기계번역에 대한 이해도가 높은 도메인 전문가들은 지표(Metric)을 추가하는 것이 Noise 병렬 말뭉치를 찾는데 도움이 될 수 있다고 대다수가 언급하였다. 현재 우리 시스템은 총 9개의 지표를 제공하고 이를 시각화 하기 위해서 각기 다른 고유한 색상으로 시각적 인코딩이 되어있다. 하지만 더 많은 지표가 추가된다면 사용자가 고유한 색상을 구별하는데 방해될 수 있기 때문에 이러한 작업은 신중해야 한다 [22]. 뿐만 아니라 지표 추가시에는 Ranking View에서 제공하고 있는 표 내에 더 많은 열이 추가됨에 따라 Scroll을 해야 하는 인터페이스를 필요로 할 수 있다. 이러한 단점을 극복하기 위해서는 사용자가 실제 관심있어 하는 몇몇개의 지표를 선택할 수 있고 그에 따라 전체적인 시각적 구성 요소가 동적으로 바뀔 수 있도록 사용자 지정 옵션을 제공해야 할 필요성이 대두된다.

신뢰성 측면

우리 시스템은 지표를 추출하기 위해 전처리 단계에서 문장을 Embedding Vector로 변환을 하며 이때 범용적으로 많이 쓰이고 있는 학습된 모델을 사용한다. 우리가 사용하는 있는 Universal Sentence Encoder는 이미 우수한 성능을 가진 것으로 알려져 있지만, 완벽하지는 않다. 특히 해당 Encoder가 특정 단어나 문자를 미리 학습하지 않았을 경우에는 Embedding Vector 자체에 의미론적 데이터가 누락됐을 수

있기에 부정확한 데이터를 포함할 수 있다. 또한 역번역(Back-Translation) 기법에 사용되는 Google Translator는 전세계에서 가장 많이 사용되는 번역 도구이지만 오역을 할 수 있기에 사용자는 지표들에 대한 신뢰성 한계를 명확히 이해하고 있어야 한다.

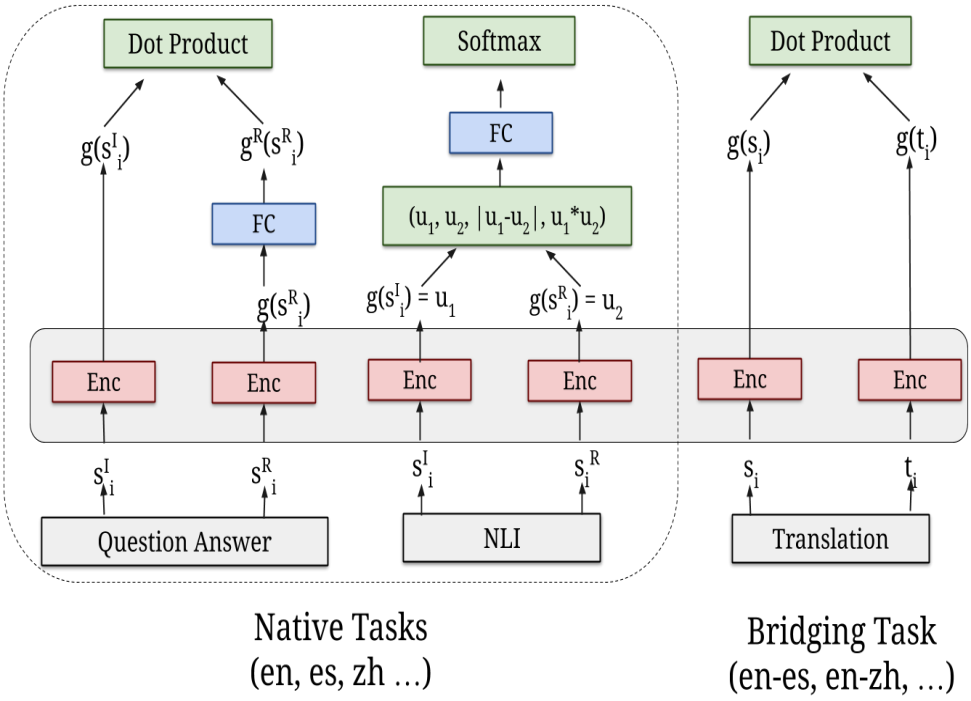

그림6. Universal Sentence Encoder의 Model Architecture

확장성 측면

일반적으로 신경망을 이용한 기계 번역 학습에서 사용하는 데이터셋인 병렬 말뭉치는 매우 큰 사이즈를 가지고 있다. 하지만 우리 시스템은 성능 문제로 인하여 데이터의 크기가 매우 커질 경우 Distribution View내에서 PCP에서 제공하는 브러싱(Brushing) 상호작용이 느려질 수 있으며, Ranking View에서 가중치를 조정할 경우도 시간적 지연 현상이 생길 수 있다. 우리는 충분한 테스트를 통해 병렬 말뭉치의 크기가 100,000까지 충분한 성능을 발휘할 수 있는 것으로

확인되었으나, 더 큰 크기의 데이터 지원을 위해서 Back-End 서버 등을 별도로 만들어서 시스템 성능을 높이는 것에 대해 검토해야 할 필요성이 있다.

제 8 장 결 론

우리 시스템은 기계 번역 학습에서 사용되고 있는 병렬 말뭉치 데이터를 정제하기 위해 사용자에게 도움이 될 수 있는 대화형 시각적 분석 시스템을 개발하였다. 우리는 기계학습(Machine Learning) 기법을 기반으로 다양한 품질 지표를 도출하였다. 사용성 연구를 통해 실제 사용자가 병렬 말뭉치 내에서 Noise 문장 쌍을 쉽게 조사하고 선별할 수 있음을 보여줌으로써 우리 시스템의 유용성과 효율성 증명하였다. 추후 사용자가 우리 시스템을 활용하여 병렬 말뭉치의 품질을 향상시키고 이를 통해 신경망을 통한 기계 번역 학습의 모델의 성능을 더욱 향상시킬 수 있을 것으로 기대한다.

참고 문헌

[1] H. Khayrallah and P. Koehn. On the impact of various types of noise on neural machine translation. arXiv preprint arXiv:1805.12282, 2018.

[2] P. Koehn, H. Khayrallah, K. Heafield, and M. L. Forcada. Findings of the WMT 2018 shared task on parallel corpus filtering. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, pp. 726–739. Association for Computational Linguistics, Belgium, Brussels, Oct. 2018. doi: 10.18653/v1/W18-6453

[3] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan,W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions, pp. 177–180, 2007

[4] B. Zhang, A. Nagesh, and K. Knight. Parallel corpus filtering via pre-trained language models. arXiv preprint arXiv:2005.06166, 2020.

[5] F. Bane and A. Zaretskaya. Selecting the best data filtering method for nmt training. In Proceedings of Machine Translation Summit XVIII: Users and Providers Track, pp. 89–97, 2021.

[7] G. Xu, Y. Ko, and J. Seo. Improving neural machine translation by filtering synthetic parallel data. Entropy, 21(12):1213, 2019.

[8] V. Dibia and C , . Demiralp. Data2vis: Automatic generation of

data visualizations using sequence-to-sequence recurrent neural networks. IEEE computer graphics and applications, 39(5):33-46, 2019.

[9] M. Rikters, M. Fishel, and O. Bojar. Visualizing neural machine translation attention and confidence. The Prague Bulletin of Mathematical Linguistics, 109(1):39, 2017.

[10] T. Munz, D. V¨¨ath, P. Kuznecov, T. Vu, and D. Weiskopf. Visualinteractive neural machine translation. In Graphics Interface 2021, 2021.

[11] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 3363-3372, 2011.

[12] D. Cer, Y. Yang, S. yi Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y.-H. Sung, B. Strope, and R. Kurzweil. Universal sentence encoder, 2018.

[13] R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. arXiv preprint arXiv:1511.06709, 2015.

[14] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311-318, 2002.

[15] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments.

In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp. 65–72, 2005.

[16] W. B. Cavnar, J. M. Trenkle, et al. N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, vol. 161175. Citeseer, 1994.

[17] J. Bok, B. Kim, and J. Seo. Augmenting parallel coordinates plots with color-coded stacked histograms. IEEE Transactions on Visualization and Computer Graphics, 2020.

[18] Y. Kim, H. Jeon, Y.-H. Kim, Y. Ki, H. Song, and J. Seo. Visualization support for multi-criteria decision making in software issue propagation. In 2021 IEEE 14th Pacific Visualization Symposium (PacificVis), pp. 81–85. IEEE, 2021.

[19] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. Lineup: Visual analysis of multi-attribute rankings. IEEE transactions on visualization and computer graphics, 19(12):2277–2286, 2013.

[20] M. Neves, A. J. Yepes, and A. N ´ ev ´ eol. The scielo corpus: a parallel corpus of scientific publications for biomedicine. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC' 16), pp. 2942–2948, 2016.

[21] J. Park, J.-P. Hong, and J.-W. Cha. Korean language resources for everyone. In Proceedings of the 30th Pacific Asia conference on language, information and computation:

Oral Papers, pp. 49–58, 2016.

[22] T. Munzner. Visualization analysis and design. CRC press, 2014.

Abstract

A Visual Analytics System for evaluating dataset of Neural Machine Translation

Sebeom Park
Computer Science and Engineering
The Graduate School
Seoul National University

The most important part of training a Neural Machine Translation model maintains good quality of parallel corpora, which are composed of pairs of different languages, Therefore, various refinement tasks have been introduced to improve the quality of parallel corpora, but there is still much room for improvement. This paper introduces a visual analysis system which helps the good quality of parallel corpora for machine translation learning. Our system provides nine different metrics in order to discover and select noise of parallel corpora. Based on our metric and visualization technics, users can find and check noise parallel corpora easily. Our system's effectiveness and usefulness are demonstrated through a qualitative user study with a total of eight users including four experts.

Keywords : NMT, data refinement, Visual Analytic System
Student Number : 2021-29094