

3rd Software Requirement Specifications for ELECTRA 기반 영화평 감성분석 System

Project Team

Team 6 (RACCOON)

Date

2020-09-15

Team Information

201611251 공민정

201611276 이규은

201611309 최지현

201612368 이지우

Table of Contents

- 1 Introduction
 - 1.1 Purpose
 - 1.2 Scope
 - 1.2.1 개발팀
 - 1.2.2 제한사항
 - 1.2.3 제품의 활용도
 - 1.2.4 개발환경
 - 1.3 Definitions, acronyms, and abbreviations
 - 1.4 References
 - 1.5 Overview
- 2 Overall Description
 - 2.1 Product perspective
 - 2.1.1 System interfaces
 - 2.1.2 User interfaces
 - 2.1.3 Hardware interfaces
 - 2.1.4 Software interfaces
 - 2.2 Product functions
 - 2.2.1 Movie Review-Input
 - 2.2.2 Pre-processing
 - 2.2.3 Training
 - 2.2.4 Show Result
 - 2.2.5 Feedback
 - 2.2.6 Save Review DB
 - 2.2.7 Refine Model
 - 2.2.8 Save Model DB

- 2.3 User characteristics
- 2.4 Constraints
- 2.5 Assumptions and Dependencies
- 2.6 Apportioning of requirements
- 3 Specific Requirements
 - 3.1 External interface
 - 3.1.1 User interface
 - 3.1.2 HW interface
 - 3.1.3 SW interface
 - 3.2 Functional requirements
 - 3.3 Nonfunctional requirements
 - 3.4 Logical database requirements
 - 3.5 System test case
- 4 Structured Analysis
 - 4.1 Architecture Diagram
 - 4.1.1 Architecture Diagram
 - 4.1.2 Component Description & Requirements
 - 4.1.3 Interface Description
 - 4.2 Model-Layer Description
 - 4.2.1 Overall Model Layer
 - 4.2.2 Description

1 Introduction

1.1 Purpose

본 문서는 'ELECTRA를 이용한 한국어 영화평 감정분석 시스템' 프로젝트를 설명하기 위한 문서이다.

1.2 Scope

1.2.1 개발팀

RACCOON(Team 6)

1.2.2 제한사항

영화평 감정분석의 결과값은 긍정적(1) 또는 부정적(0)으로만 보여진다.

특정 영화의 전반적인 영화평 감정분석 결과값을 확인하기보다는 사용자가 작성한 개별 영화평에 대해서 감정분석 결과값을 확인한다.

1.2.3 제품의 활용도

개발이 완료된 후, 사용자가 영화평을 입력하면 해당 영화평 감정분석 결과값이 긍정적(1)인지 부정적(0)인지 확인할 수 있다. 사용자는 자신이 예측한 결과값과 감정분석의 결과값이 일치/불일치하는 경우 피드백을 전달할 수 있다.

따라서 사용자가 감정분석 결과값이 궁금한 영화평에 대해 확인할 수 있는 시스템에 본 프로젝트에서 구현한 모델을 사용할 수 있다. 사용자를 통해 전달된 피드백을 통해 모델의 정확도를 개선할 수 있다.

1.2.4 개발환경

언어: Python

IDE: Jupyter notebook, Google Colab

1.3 Definitions, acronyms, and abbreviations

Name	Description
SW	Software
HW	Hardware
NN	Neural Network (인공 신경망)
ELECTRA	pre-training 기법을 적용한 language model
KoELECTRA	한국어를 지원하는 ELECTRA
KNU 감성사전	표준국어대사전을 구성하는 각 단어의 긍/부정어를 추출한 사전
NSMC	Naver Sentiment Movie Corpus (네이버 영화평)
RTD	Replaced Token Detection: ELECTRA에서 고안된 개념으로, 영화평의 일부 token을 다른 token으로 대체하여 model을 발전시키는 pre-training 방법
CLS	입력되는 영화평의 맨 앞에 붙는 token으로, 문맥 파악을 위한 구분자
FNN	FeedForward Neural Network: 연산이 Input Layer부터 Hidden Layer, Output Layer까지 순차적으로 진행되어 피드백이나 Loop를 거치지 않는 인공 신경망
bi-LSTM	양방향 LSTM (Long Short-Term Memory: RNN의 한 종류로, 장기 기억을 통해 문맥을 파악할 수 있는 인공 신경망)
User Response DB	사용자가 입력한 영화평과 그에 맞는 감정 분석 결과값을 저장하는 DB
Log DB	Model을 버전별로 관리하기 위해 저장하는 DB

1.4 References

ELECTRA: <https://openreview.net/forum?id=r1xMH1BtvB>

Character-Level Korean ELECTRA Model : <https://github.com/monologg/KoCharELECTRA>

박천음, 이창기, "BERT 기반 Variational Inference와 RNN을 이용한 한국어 영화평 감성 분석", 2019

Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning, "ELECTRA: Pretraining Text Encoders as Discriminations Rather Than Generators", 2020

IEEE Std. 830-1998

1.5 Overview

본 문서에서는 'ELECTRA를 이용한 영화평 감정분석 시스템'의 타겟 사용자에게 대한 분석과 그에 맞는 기능, 비기능적 요구사항을 서술하고, 해당 요구사항들을 만족하는 기능과 인터페이스, 시스템의 검증 및 확인을 위한 시스템 테스트 케이스를 정의한다. 또한, 시스템에 대한 이해를 돕기 위한 아키텍처 다이어그램, 모델 레이어 구성도 등을 포함한다.

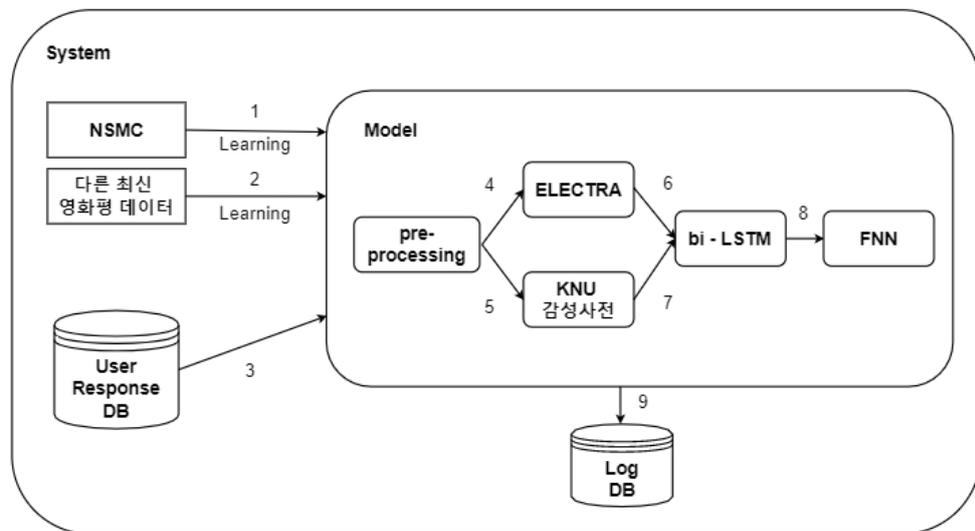
2 Overall Description

2.1 Product Perspective

본 시스템은 학습된 데이터를 기반으로 영화평 감정분석을 통해 결과값으로 영화평을 긍정 및 부정으로 분류할 수 있는 시스템이다.

사용자를 통해 영화평을 자체 구현할 UI를 통해 입력받는다. UI에 출력된 시스템 결과값과 사용자가 예측한 결과값의 일치/불일치를 사용자가 UI를 통해 선택할 수 있도록 한다. 제품의 지속적인 발전을 위해 사용자가 선택한 피드백 데이터를 DB에 저장하고, DB를 통해 업데이트되는 모델에 대한 정보도 DB에 저장해 시스템을 관리한다.

2.1.1 System Interfaces



2.1.1.1 Components

No.	Name	Definition
1	System	감정분석을 위한 전체 모델을 포함하는 시스템
2	NSMC	Naver Sentiment Movie Corpus
3	다른 최신 영화평 데이터	성능 개선을 위한 Naver 외 다른 Site의 영화평 데이터
4	User Response DB	User가 새로 입력한 response를 저장해두는 database

5	Log DB	변화하는 model을 버전별로 관리하는 database
6	Model	영화평을 학습하기 위한 model
7	pre-processing	데이터 전처리
8	ELECTRA	token(→real/fake)을 통해 학습하는 모델
9	KNU 감성사전	각 단어의 입력을 통해 긍/부정 정도 값 출력
10	bi-LSTM	영화평 전체의 문맥 분석을 통해 긍/부정 결과를 도출하는 인공신경망
11	FNN	감정분석 결과값을 위한 인공신경망

2.1.1.2 Interface

No.	Sender → Receiver	Definition
1	NSMC → Model	Model이 NSMC의 영화평을 학습한다.
2	다른 최신 영화평 데이터 → Model	Model이 NSMC 이외의 다른 영화평을 학습한다.
3	User Response DB → Model	Model이 User Response DB에 저장된 영화평을 학습한다.
4	pre-processing → ELECTRA	ELECTRA가 전처리한 word pieces로 학습한다.
5	pre-processing → KNU 감성사전	KNU 감성사전으로 word pieces의 긍/부정을 판단한다.
6	ELECTRA → bi-LSTM	학습한 word pieces들의 문치를 통해 문맥을 파악한다.
7	KNU 감성사전 → bi-LSTM	판단한 word pieces들의 문치를 통해 문맥을 파악한다.
8	bi-LSTM → FNN	문맥을 분석해 결과를 도출한다.
9	Model → Log DB	Model을 버전 별로 나누어 저장한다.

2.1.2 User interfaces

2.1.2.1 Input Window - Text



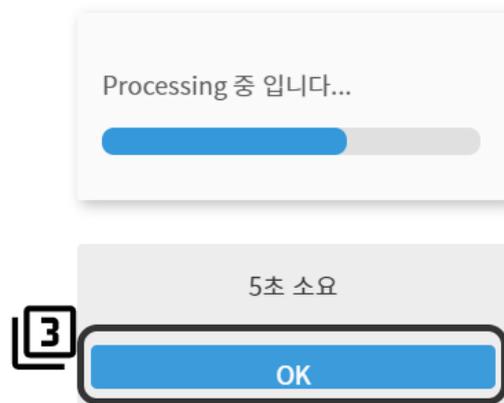
사용자가 영화평을 입력할 수 있는 Text 창

2.1.2.2 Input Window - Button



사용자가 영화평을 입력 완료했을 때 클릭하는 버튼

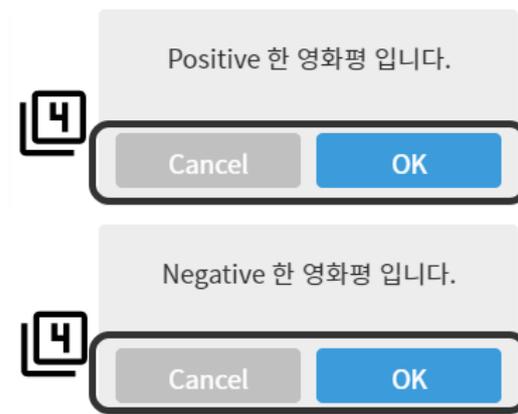
2.1.2.3 Processing Window



사용자가 입력한 영화평에 대한 결과를 보기 위해 누르는 버튼

(사용자가 입력한 영화평에 대한 감정분석이 이루어지고 있는 동안에는 상위의 창이 노출되다, 분석이 완료되면 소요시간과 결과를 보기 위해 누르는 버튼이 노출된다.)

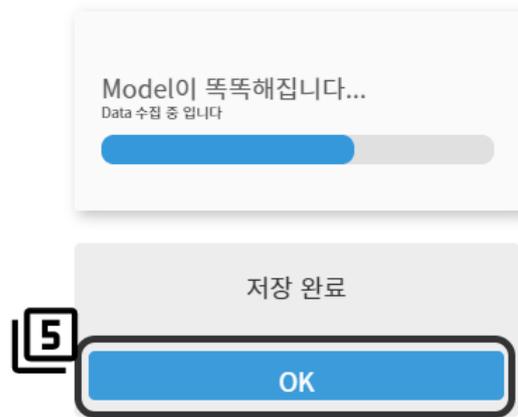
2.1.2.4 Result Window



사용자가 영화평 감정분석 결과값에 대한 피드백을 전달하는 버튼

(사용자가 예측한 결과값과 일치하면 OK, 불일치하면 Cancel을 선택한다.)

2.1.2.5 Feedback



초기 화면으로 돌아가는 버튼

(Feedback Data가 User response DB에 저장되고 있는 동안에는 상위의 창이 노출되다, 저장이 완료되면 초기화면으로 돌아가기 위한 버튼이 노출된다.)

2.1.3 Hardware Interfaces

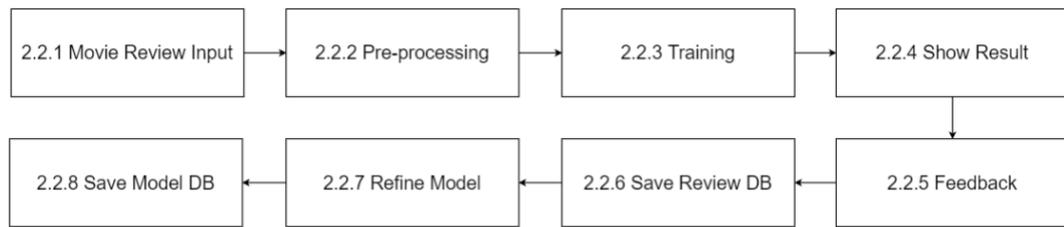
PC로만 접근이 가능하다.

키보드로 영화평을 입력 받고, 버튼은 마우스를 통해 입력된다.

2.1.4 Software interfaces

No.	Name	Version
1	torch	1.5.1
2	transformers	3.0.2
3	keras	2.3.0
4	scikitlearn	0.24
5	pandas	1.0

2.2 Product functions



2.2.1 Movie Review Input

1개의 영화평을 입력한다.

2.2.2 Pre-processing

입력된 영화평을 Pre-processing Layer에 넣어 전처리 과정을 진행한다.

2.2.3 Training

전처리 된 영화평을 Model에 넣어 긍정 및 부정의 결과값을 산출한다.

2.2.4 Show Result

영화평에 대한 긍정 및 부정의 결과값을 표시한다.

2.2.5 Feedback

영화평에 대한 결과값에 대해 사용자가 예측한 결과값과 일치/불일치하는지 선택한다.

2.2.6 Save Review DB

입력된 영화평과 피드백된 결과값을 User Response DB에 저장한다.

2.2.7 Refine Model

User Response DB에 저장된 데이터가 일정량 이상이 되면, 자동으로 Model Refine을 진행한다.

2.2.8 Save Model DB

정확도가 향상된 Model을 Log DB에 저장한다.

2.3 User characteristics

영화를 실제 감상하여 자신의 영화평을 나누고 싶은 관람객, 혹은 외부에서 수집한 영화 평의 긍/부정 결과를 알고 싶은 사람들이 영화평을 시스템에 등록한다.

컴퓨터를 사용한 경험이 있는 사람들이라면 컴퓨터 공학적 지식이 없어도 시스템 이용이 가능하다.

2.4 Constraints

영화평은 1000자 이내로 작성할 수 있다.

한국어만 가능하다.

2.5 Assumptions and Dependencies

사용자는 관람한 영화에 대해 영화평을 작성한다.

사용자는 정직한 피드백을 선택한다.

한번 입력된 영화평은 다시 입력되지 않는다.

2.6 Apportioning of requirements

시스템 구현 시 학습될 데이터는 3가지 중에서 선택될 수 있다.

- NSMC(2016년 이전의 영화평)만 사용
- 네이버 2017년부터 2019년의 영화평만 사용
- 위의 두가지 데이터 모두 사용

3가지 데이터를 각각 학습시키고, 가장 높은 정확도를 보인 모델을 최종적으로 선정한다.

3 Specific Requirements

3.1 External interface

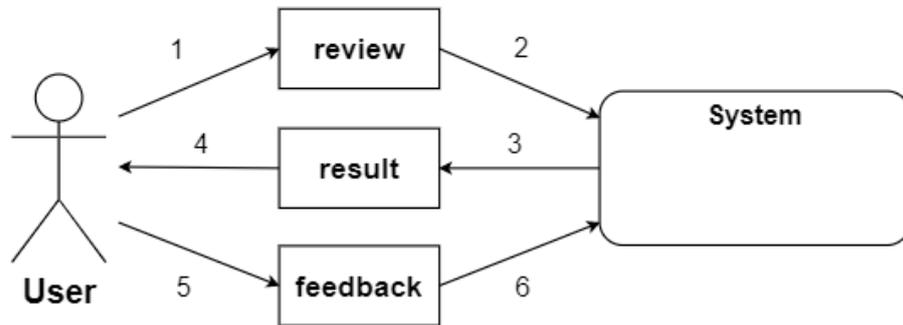
3.1.1 User interface

입력 : 키보드, 마우스

출력 : 화면

3.1.2 HW interface

3.1.3 SW interface



3.1.3.1 Components

No.	Name	Definition
1	User	시스템에 접근하는 사용자
2	review	User가 입력하는 영화평
3	result	User에게 반환하는 결과값
4	feedback	User가 result의 옳고 그름을 판단하여 입력

3.1.3.2 Interfaces

No.	Sender → Receiver	Definition
1	User → review	User가 영화평을 작성한다.

2	review → System	User가 작성한 영화평이 System으로 전달된다.
3	System → result	System이 분석한 영화평의 감정 결과값을 반환한다.
4	result → User	결과값이 User에게 전달된다.
5	User → feedback	User의 감정분석 result에 대한 feedback이 입력된다.
6	feedback → System	User의 feedback이 System으로 전달된다.

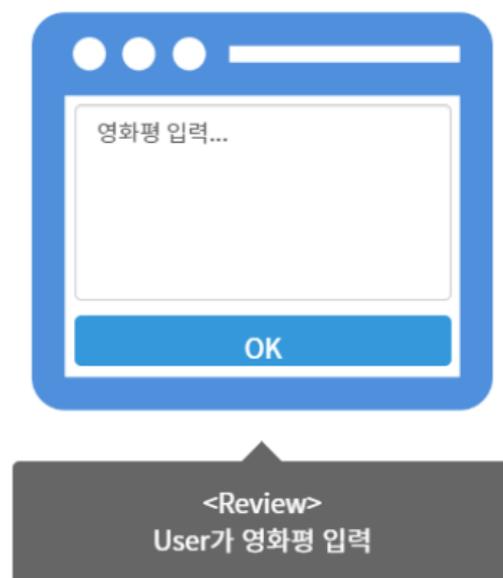
3.1.3.3 입력

Text창과 버튼

3.1.3.4 출력

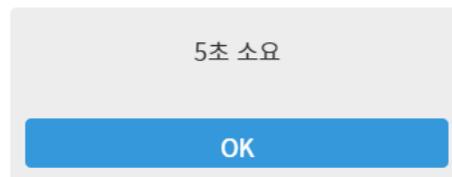
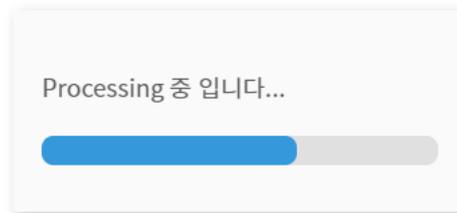
3.1.3.4.1 Review 입력

사용자가 '영화평 입력' 부분에 영화평을 입력하는 동시에 출력



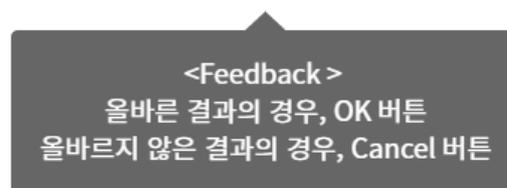
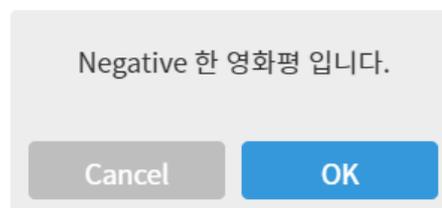
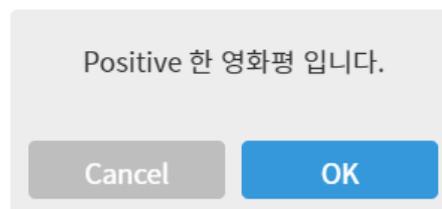
3.1.3.4.2 데이터 가공 및 감정분석 Processing

영화평 감정분석 모델을 통해 결과값을 도출하는 과정으로 과정이 끝난 뒤, 소요 시간 표시



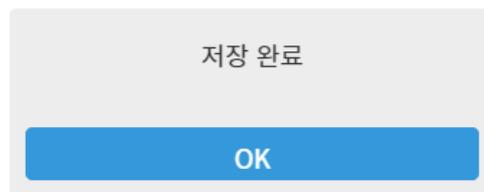
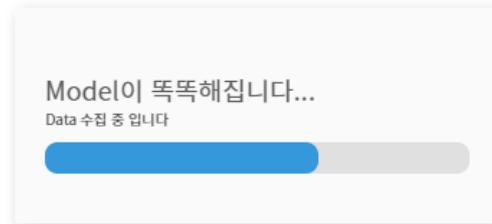
3.1.3.4.3 Feedback 선택

사용자가 예측한 결과값과 팝업창의 Positive/Negative가 일치하는 경우 'OK' 버튼 선택, 일치하지 않는 경우 'Cancel' 버튼 선택



3.1.3.4.4 DB에 저장

사용자의 피드백을 User response DB에 저장



3.2 Functional requirements

3.2.1 데이터를 가공한다.

3.2.1.1 불필요한 데이터를 제거한다.

3.2.1.1.1 이메일(ex@gmail.com)형식을 제거한다.

3.2.1.1.2 크롤링 과정에서 발생한 실제 리뷰 이외의 jQuery Hash 문자열을 삭제한다..

3.2.1.2 한글, 영어, 띄어쓰기, 일부 특수 문자 등을 제외하고 모두 제거한다.

3.2.1.2.1 한글,영어,숫자 이외의 문자를 공백으로 치환한다.

3.2.1.2.2 유니코드 이외의 이모티콘은 삭제한다.

3.2.1.3 데이터가 같은 길이가 되도록 한다.

3.2.1.3.1 tokenize된 영화평이 설정한 길이를 가지도록 padding한다.

3.2.2 모델을 구축한다.

3.2.2.1 ELECTRA와 KNU 감성사전을 이용하여 분석한다.

3.2.2.1.1 입력된 영화평을 Word Piece로 나눈다.

3.2.2.1.2 전처리를 마친 token을 ELECTRA의 Input으로 넣고, 전처리 후 형태소 분석까지 거친 token을 KNU 감성사전의 input으로 넣어 각자의 output을 산출해 둘을 엮은 vector를 만든다.

3.2.2.2 bi-LSTM을 통해 긍부정 값을 산출한다.

3.2.2.2.1 ELECTRA의 결과와 KNU 감성사전을 look up한 결과가 합쳐진 vector를 input으로 넣어 학습시킨 bi-LSTM의 결과를 통해 최종 결과를 산출한다.

3.2.2.3 RNN을 통해 최종 결과값을 산출한다.

3.2.2.3.1 ELECTRA의 CLS와 bi-LSTM의 output을 RNN의 input으로 하여 최종적인 긍/부정 결과를 얻는다.

3.2.3 감정분석 시스템 구현

3.2.3.1 사용자가 영화평을 입력한다.

3.2.3.1.1 웹을 통해 입력한다.

3.2.3.2 사용자가 입력한 영화평에 대해 감정 분석 결과를 반환한다.

3.2.3.2.1 웹을 통해 반환 결과를 알도록 한다.

3.2.3.3 사용자가 피드백을 한다.

3.2.3.3.1 감정분석 결과에 대해 피드백을 요청한다.

3.3 Nonfunctional requirements

3.3.1 85프로 이상의 정확도 달성

3.3.1.1 기존 ELECTRA 단일모델의 정확도 (Base: 85.1%, Small- 79.9%)보다 높은 정확도를 기대한다.

3.3.1.1.1 의도한 결과가 나오는지 여러 영화평을 입력해보며 확인한다.

3.3.2 짧은 응답 시간

3.3.2.1 사용자가 대기할 수 있는 응답 시간을 기대한다.

3.3.2.1.1 5초 이내에 긍/ 부정 결과를 도출하여 사용자에게 보여준다.

3.3.3 Model Quality 향상

3.3.3.1 DB를 통한 model quality 향상

3.3.3.1.1 User Response DB를 통해 User가 입력한 새로운 영화평을 저장 후 Model을 학습한다.

3.3.3.1.2 Log DB를 통해 Model을 버전별로 관리한다.

3.4 Logical database requirements

3.4.1 User response DB

3.4.1.1 Review Table

사용자의 Review를 관리한다.

Field name	Data type	Attributes	Description
seqNum	Int(10)	PK	Primary key로 자동으로 증가한다.
review	varchar(1000)		영화평
emotion	int(1)		감정분석 결과로 긍정은 1로 부정은 0으로 저장한다.

3.4.2 Log DB

History Table과 Model Table은 일대다의 관계를 갖는다.

3.4.2.1 Model Table

새롭게 개선된 모델을 관리한다.

Field name	Data type	Attributes	Description
model_id	Int(10)	PK	Primary key로 자동으로 증가한다.
model_name	varchar(10)		Model name

3.4.2.2 History Table

각 모델의 정확도를 관리한다.

Field name	Data type	Attributes	Description
history_id	Int(10)	PK	Primary key로 자동으로 증가한다.
model_id	Int(10)	FK	Model Table의 PK를 참조한다.
version	Int(10)		Model Version
precision	Int(3)		정확도

3.5 System test case

3.5.1 Pre-processing Data

Case No.	1
Name	Pre-processing Data
Description	입력된 영화평이 어절별로 나뉜다.
Input	영화평
Output	어절별로 나뉜 영화평
Success Criteria	한 영화평의 85% 이상이 어절별로 나누어진다.
Reference No.	기능 요구사항 1.1

3.5.2 Process ELECTRA

Case No.	2
Name	Process ELECTRA
Description	ELECTRA를 통해 분석 결과가 도출된다.
Input	wordpieces로 이루어진 token
Output	CLS를 포함한 replace된/되지 않은 token
Success Criteria	긍/부정 값의 정확도가 85% 이상이다.
Reference No.	기능 요구사항 1.2

3.5.3 Process KNU Dictionary

Case No.	3
Name	Process KNU Dictionary
Description	KNU 감성사전을 이용해 분석 결과가 도출된다.
Input	wordpieces로 이루어진 token
Output	5-degree로 나누어진 wordpieces의 긍/부정의 정도
Success Criteria	각 wordpiece의 긍/부정 값이 감성 사전에 정의된 대로 도출된다.
Reference No.	기능 요구사항 1.3

3.5.4 Process bi-LSTM Output

Case No.	4
Name	Process bi-LSTM Output
Description	bi-LSTM을 통해 결과가 도출된다.
Input	ELECTA output과 KNU 감성사전 output의 mapping vector

Output	금/부정 결과
Success Criteria	금/부정 값이 올바르게 도출된다.
Reference No.	기능 요구사항 1.4

3.5.5 Process Input Data

Case No.	5
Name	Process Input Data
Description	임의의 영화평 입력 시, 감정 분석 결과가 반환된다.
Input	영화평
Output	금/부정 결과
Success Criteria	감정 분석 결과의 정확도가 85% 이상이다.
Reference No.	기능 요구사항 2.1

3.5.6 High Accuracy

Case No.	6
Name	High Accuracy
Description	ELECTRA 단일 모델보다 높은 정확도를 달성한다.
Input	영화평
Output	금/부정 결과
Success Criteria	모델 정확도가 85% 이상인지 확인한다.
Reference No.	비기능 요구사항 1.1

3.5.7 Short Duration

Case No.	7
-----------------	---

Name	Short Duration
Description	사용자가 영화평을 입력한 후, 짧은 시간 내에 결과가 도출된다.
Input	영화평
Output	긍/부정 결과
Success Criteria	5초 이내에 결과가 도출된다.
Reference No.	비기능 요구사항 2.1

3.5.8 Additional Learning

Case No.	8
Name	Additional Learning
Description	User Response DB에 저장된 영화평으로 추가 학습을 한다.
Input	추가적인 영화평 데이터 셋
Output	추가 학습이 완료된 모델
Success Criteria	추가 학습을 통해 Model의 정확도를 향상시킨다.
Reference No.	비기능 요구사항 3.1

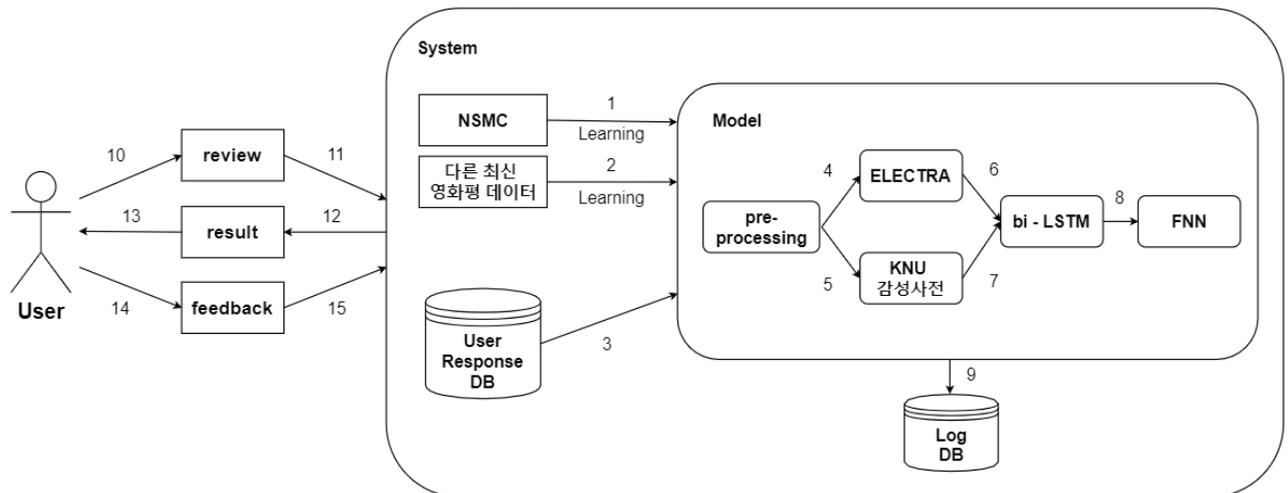
3.5.9 Model Version Control

Case No.	9
Name	Model Version Control
Description	Log DB에 각 Model을 버전별로 저장한다.
Input	현 모델
Output	버전별로 저장된 모델
Success Criteria	Log DB에 각 Model을 버전별로 저장한다.
Reference No.	비기능 요구사항 3.2

4 Structured Analysis

4.1 Architecture Diagram

4.1.1 Architecture Diagram



4.1.2 Component Description & Requirements

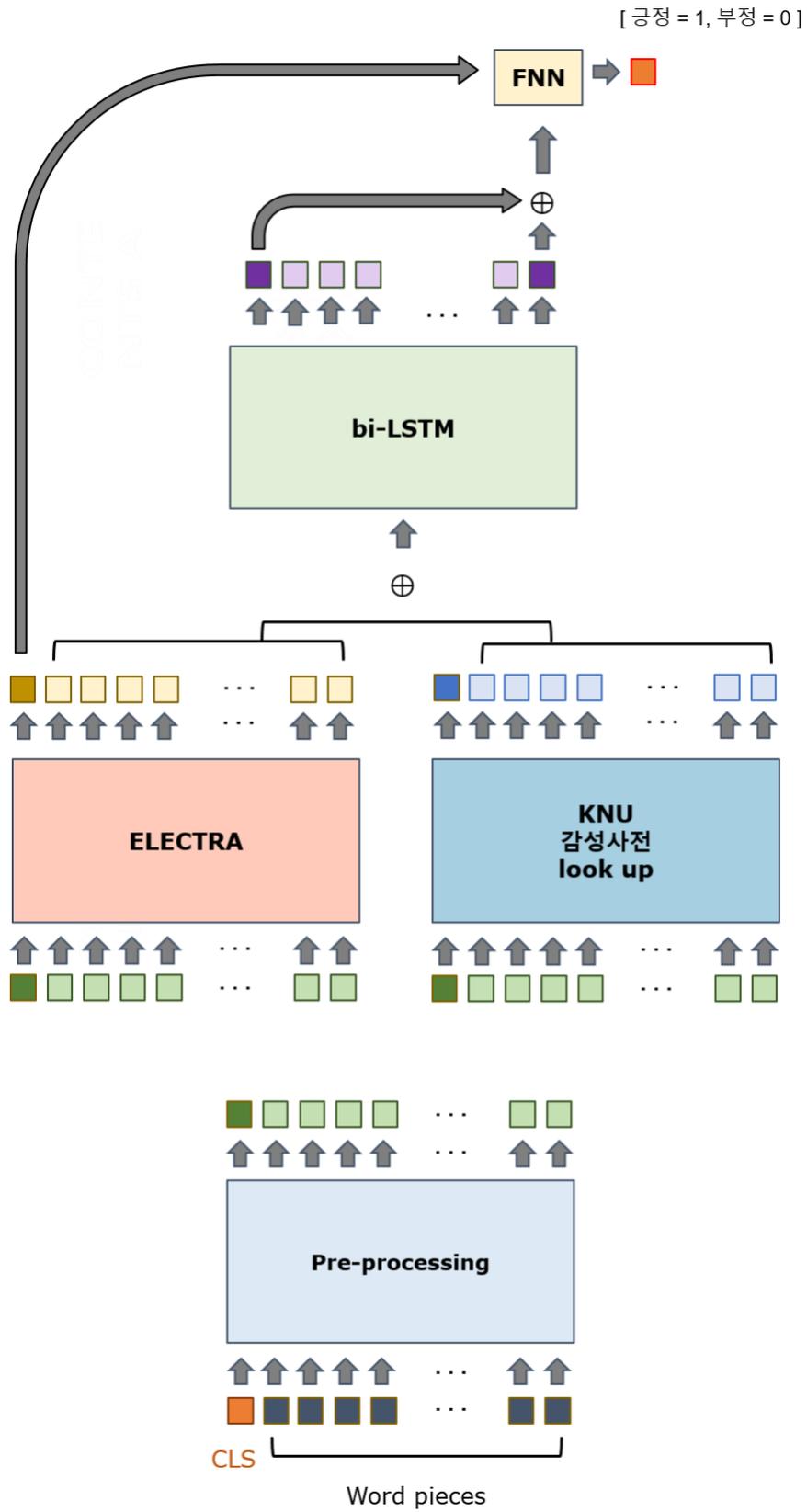
No.	Name	Definition	Requirement No.
1	System	감정분석을 위한 전체 모델을 포함하는 시스템	FR 1 / FR 2 / FR3
2	NSMC	Naver Sentiment Movie Corpus	FR 1 / FR 2
3	다른 최신 영화평 데이터	성능 개선을 위한 Naver 외 다른 Site의 영화평 데이터	FR 1 / FR 2 / NFR 1.1.1
4	User Response DB	User가 새로 입력한 response를 저장해두는 database	NFR 3.1
5	Log DB	변화하는 model을 버전별로 관리하는 database	NFR 3.2
6	Model	영화평을 학습하기 위한 model	FR 2
7	pre-processing	데이터 전처리	FR 1
8	ELECTRA	token(→real/fake)을 통해 학습하는 모델	FR 2.2
9	KNU 감성사전	각 단어의 입력을 통해 금/부정 정도값 출력	FR 2.3
10	bi-LSTM	영화평 전체의 문맥 분석을 통해 금/부정 결과를 도출하는 인공지능망	FR 2.3
11	FNN	감정분석 결과값을 위한 인공지능망	FR 2.4
12	User	시스템에 접근하는 사용자	NFR 1.1 / NFR 2.1
13	review	User가 입력하는 영화평	NFR 3.1
14	result	User에게 반환하는 결과값	NFR 1.1 / NFR 2.1
15	feedback	User가 result의 옳고 그름을 판단하여 입력	NFR 3.1

4.1.3 Interface Description

No.	Sender → Receiver	Definition
1	NSMC → Model	Model이 NSMC의 영화평을 학습한다.
2	다른 최신 영화평 데이터 → Model	Model이 NSMC 이외의 다른 영화평을 학습한다.
3	User Response DB → Model	Model이 User Response DB에 저장된 영화평을 학습한다.
4	pre-processing → ELECTRA	ELECTRA가 전처리한 word pieces로 학습한다.
5	pre-processing → KNU 감성사전	KNU 감성사전으로 word pieces의 긍/부정을 판단한다.
6	ELECTRA → bi-LSTM	학습한 word pieces들의 뭉치를 통해 문맥을 파악한다.
7	KNU 감성사전 → bi-LSTM	판단한 word pieces들의 뭉치를 통해 문맥을 파악한다.
8	bi-LSTM → FNN	문맥을 분석해 결과를 도출한다.
9	Model → Log DB	Model을 버전 별로 나누어 저장한다.
10	User → review	User가 영화평을 작성한다.
11	review → System	User가 작성한 영화평이 System으로 전달된다.
12	System → result	System이 분석한 영화평의 감정 결과값을 반환한다.
13	result → User	결과값이 User에게 전달된다.
14	User → feedback	User의 감정분석 result에 대한 feedback이 입력된다.
15	feedback → System	User의 feedback이 System으로 전달된다.

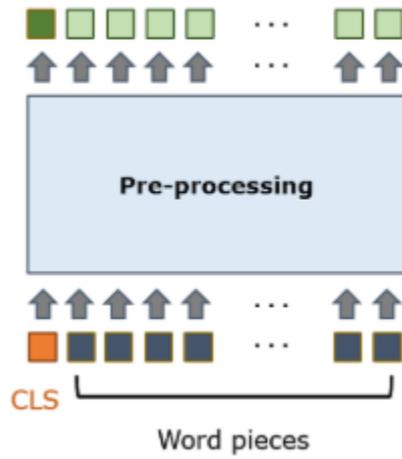
4.2 Model- Layer Description

4.2.1 Overall Model Layer



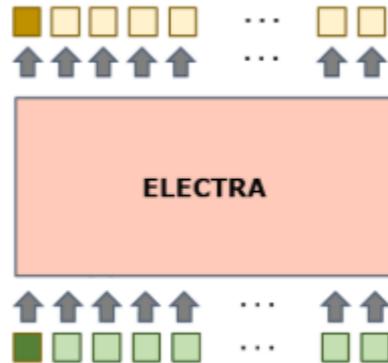
4.2.2 Description

4.2.2.1 Pre-processing



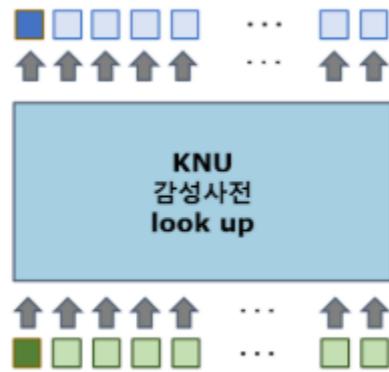
한개 이상의 문장으로 이루어진 영화평을 ELECTRA 혹은 KNU 감성사전에 입력하기 적절한 형태로 pre-processing한다.

4.2.2.2 ELECTRA



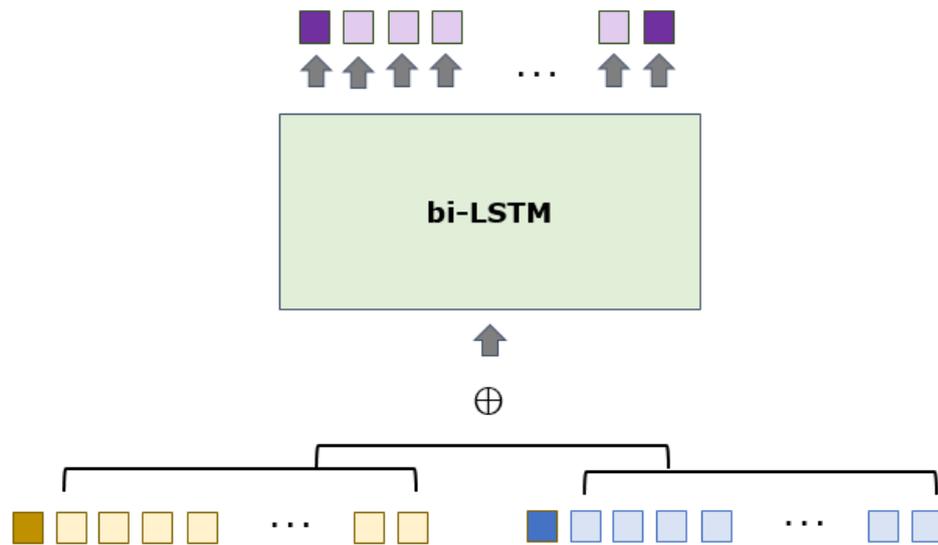
전처리한 문장의 token을 ELECTRA에 학습시킨다.

4.2.2.3 KNU 감성사전 look up



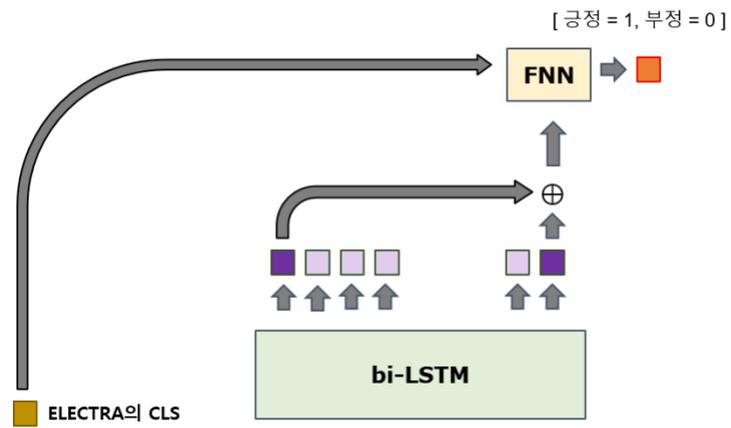
전처리한 문장의 token을 감성사전에서 look up한다.

4.2.2.4 bi-LSTM



ELECTRA의 output과 KNU 감성사전의 output의 mapping vector가 bi-LSTM의 input으로 입력되면, bi-LSTM에서 output이 산출된다.

4.2.2.5 FNN



ELECTRA의 CLS와 bi-LSTM의 output이 FNN의 input이 되어, FNN으로부터 최종적으로 금/부정 결과값이 산출된다.