감정분석 및 텍스트랭크를 이용한 영화 대표 리뷰 추출과 평점 부여 서비스

이요한 · 이정우 · 우인준 · 남상혁 · 이경규
(42 Seoul)

요 약

대부분의 사람들은 리뷰와 평점만 보고 어떤 영화를 볼 지 선택한다. 하지만 기존 시스템 상 영화와 관계없는 리뷰를 구별하기 어렵고, 평점 조작이 용이하여 예비 시청자들의 영화 선택에 방해가 될 수 있다. 본 글에서는 사용자가 작성한 리뷰의 감정분석과 평점 예측을 위한 Hard Parameter sharing 구조의 멀티태스크와 전이학습 기반의 모델을 제안하며 대표성을 띄는 리뷰를 추출하기 위해 텍스트랭크 알고리즘을 활용한다. 이에따라 '감정분석 및 텍스트랭크를 이용한 영화 대표 리뷰 추출과 평점 부여 서비스'는 영화에 대한 보다 객관적인 지표를 제공하고, 마케팅 및 영화추천 알고리즘 등에서 활용될 것으로 기대한다.

1. 서론: 리뷰와 평점의 괴리

바쁜 현대인들은 점차 '읽기' 행위를 포기하고 있다. 대신 대체로 신문의 헤드라인이나 인터넷 상의 요약문만 보고, 글을 읽을지 여부를 판단한다. 이렇게 요약된 정보로만 전체적인 판단을 하는 습성은 영화의 '한 줄 평'과 '평점'에서도 작용한다. 즉 리뷰와 평점이 사용자들의 감정을 대변하는 것뿐만 아니라 개개인들의 취향에 맞는 상품과 서비스를 선택할 수 있는 척도로 기능하는 것이다. 특히 영화는 시청에 오랜 시간을 필요로 하기 때문에, 예비 시청자들은 상대적으로 평점이 높은 영화 혹은 긍정적인 평가가 존재하는 영화들에 우선순위를 둘 것이며 다양한 경쟁작들이 존재하는 영화시장에서 평점이 낮고 부정적인 평이 많은 영화는 고려대상이 되지 않을 것이다.
이처럼 리뷰와 평점은 각각 한 영화를 대변하는 지표로써 사용되고 있지만, 한 영화에 대한 객관적인 지표라고는 할 수 없다. 단 한 줄의 리뷰가 영화의 본질을 대표할 수 없는 것은 물론이고, 리뷰작성 시 그 내용과 별개로 임의의 평점을 매길 수 있는 현 시스템 상 손쉽게 평점 조작이 가능하기 때문이다.

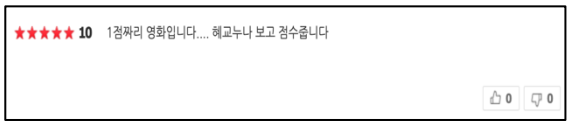

그림 1. 리뷰내용과 평점의 괴리

가령 그림 1의 경우, 리뷰 작성자는 영화가 평점 1점에 가깝다고 판단하였으나, 배우에 대한 개인적인 감정으로 영화에 임의의 평점 10점을 주어 리뷰와 평점 간에 큰 괴리가 발생하였다. 이런 상황에서 '전체 평점' 등 단순히 평점으로 구성된 지표로 영화를 판단한다면 영화의 본질과 다른 해석이 우려된다.
이렇듯 현대인들의 요약된 정보만을 읽는 습성을 반영하였을 때, 영화의 본질과 다른 해석이 가능한 리뷰와 평점 시스템은 사용자가 잘못된 기대감으로 영화를 소비할 수 있다는 것을 암시한다. 따라서 본 글은 이러한 문제에 착안하여 대표성을 띠는 리뷰를 추출하고 평점데이터를 보다 객관적인 지표로서 사용할 수 있는 방법을 제안하며, 리뷰와 평점 간의 괴리를 해결할 수 있는 서비스를 구축하고자 했다.

2. 본론: 시스템 설계

2.1 멀티태스크와 전이학습 기반의 Movie-BERT

본 서비스는 다양한 과제를 수행하기 위해 Multi-task(이하 멀티태스크) [1] 와 전이학습(Transfer Learning) 기반의 Movie-BERT 시스템을 제안한다. Movie-BERT는 그림 2와 같이 한 가지 모델로 감정분석을 통한 영화 리뷰의 긍·부정 분류와 평점 부여 과제를 수행한다.

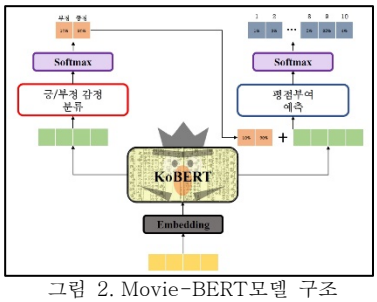

그림 2. Movie-BERT모델 구조

2.1.1 데이터 수집 및 언어모델

한국어 감정분석을 위해 긍·부정 정보가 들어있는 말뭉치는 시중에 거의 존재하지 않기 때문에, 이진 감정 분류의 안정적인 성능 보장을 위해 기존에 알려진 Naver Sentiment Movie Corpus v1.0(이하 NSMC) [2] 를 이용했으며 두 번째 과제인 평점 부여를 위해 네이버 영화 사이트에서 리뷰와 평점 데이터를 수집했다. 해당 사이트의 영화 평점들은 대부분 높게 형성되어 있고, 부정적인 댓글보다 긍정적인 댓글들이 상대적으로 많이 존재한다. 이러한 점을 고려하여 대조군의 수량을 맞추기 위해 평점

이 6~7점이며 5,000개 이상의 리뷰가 존재하는 영화 29개를 선정하였고, 통합적인 전처리 후 리뷰와 평점으로 이루어진 총 16만 개의 원시말뭉치 [3]를 구축했다. 최종적으로 사용한 언어모델은 SKT Brain의 KoBERT [4] 로, 구글에서 개발한 BERT [5] 기반의 언어모델이다. BERT는 다중언어모델로 안정적인 성능을 보장한다는 장점이 있지만, 한국어의 불규칙한 언어 변화적 특성 때문에 성능의 한계가 있었다. 이러한 문제를 해결하기 위해 대규모 한국어 위키와 뉴스로 사전학습(Pre-trained)된 KoBERT를 사용하였으며, 수집한 데이터로 미세조정(fine-tuning)하며 Movie-BERT의 학습을 진행하였다.

2.1.2 Multi-task & Transfer Learning

컴퓨터 자원의 효율과 추론 시간을 줄이기 위해 한 가지 모델이 여러 과제를 수행하는 멀티태스크 기법이 많이 사용되며, 두 가지 과제를 수행해야 하는 Movie-BERT 또한 Hard Parameter sharing 구조의 멀티태스크 기법이 적용되었다. 우선 긍·부정 과제(그림 3)을 학습한 후, 전이학습을 통해 평점부여 과제(그림 4)를 학습시켰다.

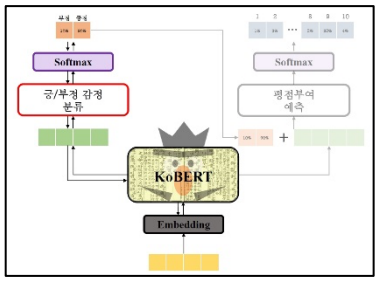

그림 3. [과제 1] 분류모델 학습 과정

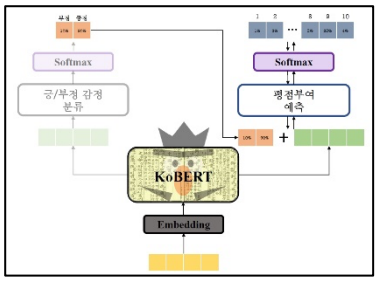

그림 4. [과제 2] 예측모델 학습 과정

2.1.3 실험 및 결과

2.1.3.1 긍정·부정 분류

NSMC의 훈련데이터(Trainset)와 평가데이터(Testset)를 사용하여 학습을 진행했다. 평가데이터로 측정한 결과 24epochs로 89.59%의 정확도를 얻을 수 있었다. 감정분류의 성능을 유지하기 위해 신경망의 매개변수(Network Parameter)들을 고정함으로써 [과제 2] 학습에 의한 성능 저하를 방지했다.

2.1.3.2 평점 예측

서로 양극의 성향인 긍·부정을 분류하는 것보다, 오직 문장만으로 1점부터 10점까지의 평점을 파악하는 것은 상당한 난이도가 요구되었다. 또한 정확한 정답을 맞추게 하는 평가 방식으로는 높은 정확도를 얻기 어려웠다. 따라서 예측된 평점이 그림 5와 같이 부정·중립·긍정 구간에 해당하면 리뷰에서 나타난 감정과 동일한 경향성을 나타낸다는 전제를 두었고, 네 개의 실험을 통해 Movie-BERT의 정확도를 측정했다. 이 때 수집한 총 16만 개의 데이터 중 14만개를 학습데이터로, 2만개를 평가데이터로 활용하였다.

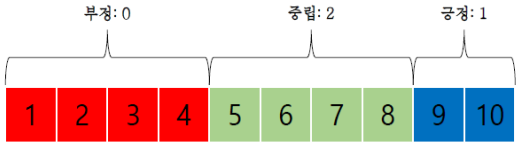

그림 5. 평점 별 감정 구간

- 실험 1. 정확한 평점을 얼마나 잘 예측하는가.
- 실험 2. 1~4점은 부정, 5~8점은 중립, 9~10점을 긍정의 성향으로 분류, 실제 성향과 얼마나 일치하는가.
- 실험 3. 각 성향의 경계점(4, 5, 8, 9점)을 제거한 후, 실제 성향과 얼마나 일치하는가.
- 실험 4. 1~4점인 부정과 9~10점인 긍정에 해당하는 리뷰만을 대상으로 실제 성향과 얼마나 일치하는가.

표 1의 결과를 보면, 완벽한 평점을 예측하는 것은 매우 어렵지만, 감정 구간을 예측한 정확도는 60%가 넘는 것을 확인할 수 있었다. 특히, 긍정/부정에 해당하는 문장에 대해서는 87.33%로 감정에 어울리는 평점이 할당되는 것을 볼 수 있었다.

	실험 1 (Acc.)	실험 2 (Acc.)	실험 3 (Acc.)	실험 4 (Acc.)
Movie-BERT	21	60	65	87.33

표 1. 감정 예측 평가

2.1.3.3. 감정분석과 평점 불일치 개선

불일치율은 서비스를 이용하는 사용자의 신뢰감을 떨어트리는 문제를 야기할 수 있다. Movie-BERT의 [과제 1]에서 부정 리뷰로 분류했지만, [과제 2]에서 5 이상의 평점이 예측되는 등, 서로 상관관계가 있는 두 가지 과제를 수행하면서 분석 결과가 일치하지 않는 문제가 발생했다. 일차적으로 모델이 100%의 정확도를 가지지 못한다는 것과 각 과제별로 서로 다른 파라미터를 가지기 때문이다. 불일치 문제를 완화시키기 위해 보조 값(hint)을 활용하였는데, 감정 분류를 수행하는 softmax 함수의 산출값을 평점 부여 과제의 입력값과 결합하여 사용했다. 세 개의 실험을 통해 모델의 개선정도를 확인할 수 있었다.

- 실험 1. 수집한 전체 데이터 전부를 사용하여 감정 분류와 평점 예측간 불일치율 평가.
- 실험 2. 수집한 데이터 중 긍·부정에 속하는 1~4점, 9~10점 데이터만 사용하여 감정 분류와 평점

예측간 불일치율 평가.
- 실험 3. NSMC의 Testset을 이용하여 감정 분류와 평점 예측간 불일치율 평가.

표 2와 같이 보조 값을 사용한 모델과 사용하지 않은 모델을 비교했을 때, 미세하지만 1~2 % 정도의 개선 효과가 있음을 확인할 수 있었다.

	실험 1 (discrepancy rate)	실험 2 (discrepancy rate)	실험 3 (discrepancy rate)
No hint	14.02	10.22	14
Hint	12.86	9.1	12

표 2. 감정 분류와 평점 예측간의 불일치율 평가

2.2 대표 문장 및 키워드 추출

Textrank(이하 텍스트랭크) [6] 는 문서 집합을 추출적요약(Extractive summarization)하는 대표적인 방법으로, 문장그래프를 구축한 뒤 구글이 제안한 Pagerank(이하 페이지랭크) 를 [7] 이용하여 키워드와 핵심문장을 선택한다.

$$sim(s_0, s_1) = \frac{|\{w_k \mid w_k \in S_1 \,\&\, w_k \in S_2\}|}{\log|S_1| + \log|S_2|}$$

수식 1. 문장 간 유사도 계산 함수 [8]

$$PR(u) = c \times \sum_{v \in B_u} \frac{PR(v)}{N_v} + (1-c) \times \frac{1}{N}$$

수식 2. 그래프 랭킹 계산 함수

우선 수집한 리뷰들에 대하여 형태소분석을 진행한 후 리뷰 간 유사도를 측정한다. 유사도는 수식 1과 같이 두 문장에 공통으로 등장 한 단어의 개수를 두 문장 단어 개수의 총합으로 나누어 계산한다. 이 때 사용하는 단어집합의 품사를 명사, 형용사, 동사로 제한하는데, '은/는', '이/가' 등 의미를 가지지 않은 품사를 포함시키면 다른 단어와의 유사도가 압도적으로 높게 나타나기 때문이다.

단어	Rank
영화/NNG	76.5
보/VV	64.2
없/VA	56.8
것/NNB	53.7
하/VV	52.7
수/NNB	36.7
년/NNBC	35.9
좋/VA	35.2
같/VA	34.6
있/VV	33.1

표 3. 모든 품사 대상

단어	Rank
영화/NNG	74.9
감독/NNG	24.3
사람/NNG	23.6
때/NNG	23.4
생각/NNG	20.7
스토리/NNG	20.1
말/NNG	18.6
작품/NNG	18.4
장면/NNG	18.2
사랑/NNG	16.8

표 4. 명사 대상

표 1은 모든 품사 대상으로 키워드를 추출했고, 표 2는 명사만을 대상으로 키워드를 추출한 결과이다. 명사로 추출된 단어들에 비해 표 1에서는 '보', '것', '하' 등 의미를 지니지 않은 형태소들이 많이 포함된 것을 확인할 수 있다. 최종적으로 유사도가 계산된 문장그래프를 페이지랭크로 보내고, 수식 2의 계산을 거쳐 리뷰들의 랭크를 계산하게 된다. 그림 3과 그림 4의 워드클라우드에는 주로 '연기', '현실', '배우'와 같은 단어들이 대거 포함되어 있고, 이 키워드들이 많이 포함된 문장이 그림 5와 그림 6에서 대표 리뷰로 추출되는 것을 확인할 수 있다.

그림 3. 영화 특별시민 긍정 워드클라우드

그림 4. 영화 특별시민 부정 워드클라우드

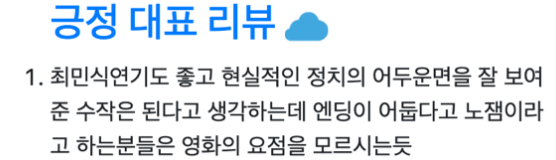

그림 5. 영화 특별시민 긍정 대표 리뷰

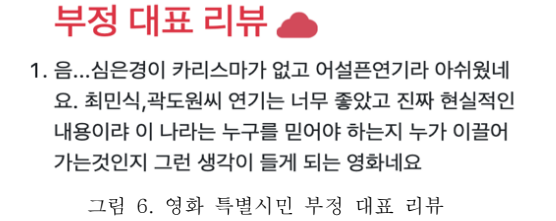

그림 6. 영화 특별시민 부정 대표 리뷰

2.3 서비스 설계

'감정분석 및 텍스트랭크를 이용한 영화 대표 리뷰 추출과 평점 부여 서비스'는 이용자로 하여금 리뷰와 평점을 지표삼아 왜곡없이 영화를 판단할 수 있도록 돕는 서비스이다. 이를 위해 1) 선택한 영화에 대해 긍·부정 대표 리뷰 각 10건을 확인하는 '리뷰큐레이션 기능'과 2) 리뷰와 별도로 평점을 매길 수는 없지만, 리뷰를 남겼을 때 문장에서 추출한 긍·부정 비율 및 평점을 확인하는 '자동 평점 부여 기능'을 이용할 수 있도록 설계하였다.

2.3.1. Front-End

Front-End(이하 프론트엔드)는 영화 목록 페이지와 리뷰분석 페이지로 구성되어 있다. 영화를 선택하면

리뷰분석 페이지로 이동하고, 긍·부정 대표 리뷰 각 10건씩과 워드클라우드, 수집 데이터의 기존평점과 예측평점을 확인할 수 있다. HTML과 CSS를 이용하여 디자인했다.

2.3.2. Back-End

Back-End(이하 백엔드)는 그림 7과 같이 사용자의 요청을 받아 응답하는 서버와 리뷰 감정분석 및 평점부여 API 서버를 따로 구성하여 프론트엔드와 연결시켰으며, 웹 서버는 AWS의 EC2 Instance를, API 서버는 Goorm IDE의 컨테이너를 사용했다. 모델의 예측을 빠르게 진행하기 위해 RAM 4GB의 CPU 성능이 높은 컨테이너를 사용했다.

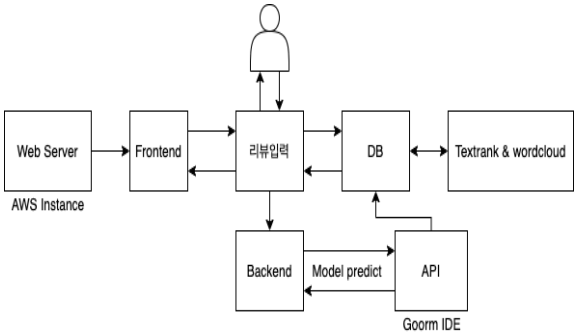

그림 7. 서비스 구조도

1,000건의 데이터를 수집하고 텍스트랭크를 적용하는 데 약 10분이 소요되는 점을 감안하여 사전에 대표 리뷰와 워드클라우드를 데이터베이스에 저장했다. 한편 사용자가 페이지에 리뷰를 입력할 경우, API 서버에서 해당 리뷰의 감정분석 결과와 예상평점을 반환한다. 이 때 반환값을 응답화면 렌더링에 활용함과 동시에 데이터베이스에 저장시킨다.

3. 결과

실제 서비스의 구현 [9] 화면은 그림 8을 통해 확인할 수 있다. 영화 '마약왕'은 네이버에 등재되어 있는 평점이 6.33인 반면에 본 시스템의 예측평점은 4.85로, 약 1.5가량 낮게 측정된 것을 볼 수 있다. 이는 부정적인 댓글들의 극성을 평점이 정확하게 담아내지 못한 것이라고 해석할 수 있다. 또한 "이 영화 정말 재미없는 듯, 다시는 보고 싶지 않은 영화."라는 리뷰를 입력했을 때 80% 확률로 부정이며, 예측 평점은 3점으로 리뷰 작성자의 마음을 평점 정보에 반영하고 있음을 확인할 수 있다.

4. 결론

4.1 활용 및 발전 가능성

이렇듯 본 서비스를 이용하게 되면 실제 리뷰에 더욱 근접한 평점 정보를 바탕으로 소비자의 영화 선택에 도움을 줄 수 있을 것이라 기대한다. 또한 감정분석 및 평점예측 API는 확장성을 고려하여 독립된 서버로 구축했기 때문에 외부에서 사용이 가능하다. 기업은 이것을 영화 홍보 및 감정 키워드를

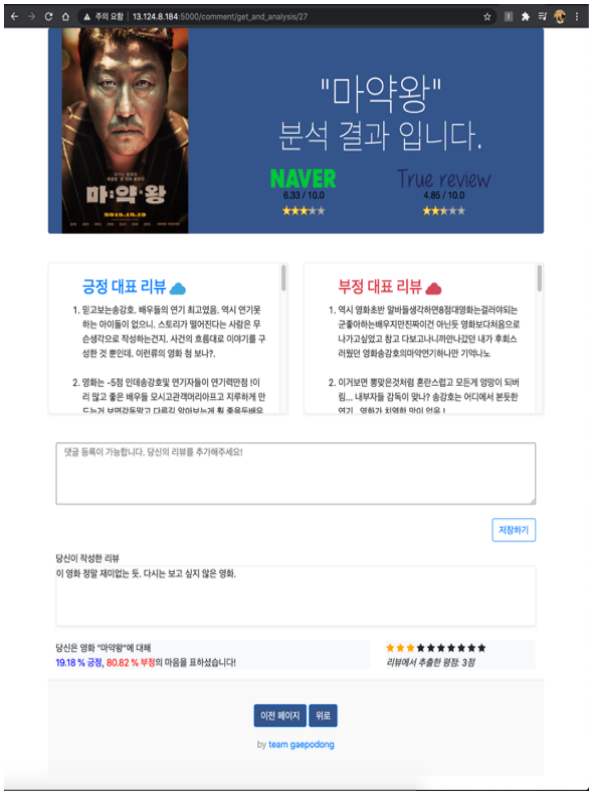

그림 8. 서비스 예제

통한 트렌드 분석의 수단으로 사용할 수 있을 것이다. 추후 세분화된 감정별 수치가 레이블링(Labeling)된 말뭉치가 구축되고, 국립국어원이나 공공데이터 포털로부터 제공받을 수 있다면, 단순 긍·부정보다 다양한 감정에 따른 유의미한 정보들을 제공할 수 있을 것이다. 감정에 따른 필터링 기능을 추가하여 보다 세부적인 조건들로 소비자들에게 맞춤형 영화를 추천해주는 서비스로 응용될 수 있을 것이며, 기업에게는 이러한 정보들을 감정 마케팅에 활용할 수 있을 것으로 예상한다.

4.2 한계 및 향후 연구

그림 9의 '진 경기에서도 노력하는 저 태도 자체가 월클이구나 싶다'라는 댓글은 스포츠 경기에서 진 선수를 칭찬하는 긍정적인 댓글이지만 본 시스템은 부정적인 댓글로 예측하고 6점의 평점을 부여했다.

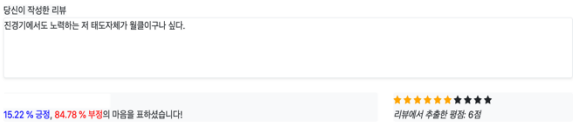

그림 9. 스포츠 댓글

이처럼 영화 외의 댓글을 제대로 분석해내지 못하는 이유는 1) Movie-BERT를 미세조정하며 활용한 데이터의 도메인이 영화에 편중되어 있기 때문이고, 2) KoBERT의 사전훈련에 사용된 데이터의 규모가 매우 작다는 것에서 기인할 수 있다. KoBERT의 사전훈련에 사용된 데이터의 크기는 25 Million으로, 여러 분야의 벤치마크에서 최첨단 성능을 달성한

GPT-3 [10] 에는 300 Billion, T5 [11] 에는 1,000 Billion 규모의 학습데이터가 사용되었다는 것에 비하면 터무니없이 적은 양이다.
이러한 문제를 해결하기 위해서는, 도메인과 분야별로 자연어 처리 과제에 적합하게 레이블링된 말뭉치들을 보다 더 많이 확보할 필요가 있다. 사용자 정의(Manual Handling)에 의한 데이터셋 구축은 매우 많은 시간과 비용을 필요로 하며, 작업 효율이 좋지 않기 때문에 대규모 학습 데이터를 필요로 하는 딥러닝에 적합하지 않다. 따라서 향후 연구로 Movie-BERT의 기존 두 가지 과제에 자동 레이블링(Auto-Labeling) 과제를 추가하여 세 가지 과제를 수행하는 멀티태스크 모델을 만들 수 있을 것이다. 자동 레이블링을 활용하면 원시 말뭉치들을 대상으로 해당 과제에 적합하게 레이블링 된 데이터를 빠른 시간에 얻을 수 있을 수 있고, 더욱 다양한 도메인에서 균일한 성능을 보장할 수 있을 것으로 예상한다.
또한 대규모 말뭉치를 학습한 사전훈련 모델을 만들 필요가 있다. 그림 10은 T5의 사전훈련 시 데이터 규모에 따른 성능 추이이다. 대체로 더 많은 데이터를 사용한 한 번의 훈련이 보다 적은 데이터를 반복하여 훈련시키는 것보다 높은 성능을 보인다.

Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★Full data set	0	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
2^{29}	64	**82.87**	**19.19**	**80.97**	**72.03**	**26.83**	**39.74**	**27.63**
2^{27}	256	82.62	**19.20**	79.78	69.97	**27.02**	**39.71**	27.33
2^{25}	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81

그림 10. T5 사전훈련 데이터 규모 별 성능비교

사전훈련 모델을 통한 미세조정과 전이학습을 위해 가장 중요한 것은 레이블링 되지 않은 원시 말뭉치의 크기이며, 데이터의 품질보다는 다양성이 더 중요하다고 한다. (Raffel, 2019) 따라서 다양한 말뭉치들을 통합시켜 대규모의 사전훈련용 말뭉치를 구축하는 것이 향후 연구 과제가 될 수 있으며, 이 과정에서 국립국어원 모두의 말뭉치를 활용할 수 있을 것이다. 모두의 말뭉치는 원시 말뭉치들을 분야에 맞게 분류해 놓았기 때문에, 제공되는 데이터들을 통합하여 학습데이터로 사용하게 된다면 기존 사전훈련 KoBERT보다 큰 규모의 말뭉치로 학습되어 성능이 향상된 사전훈련 모델을 만들 수 있을 것으로 기대한다.

5. 참고문헌

[1] Ruder, S. (2017). An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098.

[2] https://github.com/e9t/nsmc

[3] https://github.com/Gaepodong/Your-True-Review/tree/master/data/model

[4] https://github.com/SKTBrain/KoBERT

[5] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[6] Rada Mihalcea, Paul Tarau (2004). TextrRank: Bringin Order into Texts. Association for Computational Linguistics. 404-411.

[7] Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry (1999) The PageRank Citation Ranking: Bringing Order to the Web. Technical Report. Stanford InfoLab.

[8] https://lovit.github.io/nlp/2019/04/30/textrank

[9] http://13.124.8.184:5000/movies/list/

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Agarwal, S. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[11] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683.