Sentence BERT를 활용한

기업 SWOT 분석 자동화 연구

오케스트로 주식회사

정상현

※ 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00147, 멀티-하이브리드 SaaS 솔루션 통합관리 플랫폼 기술 개발)

Chapter 01

Introduction

Introduction

❖ 클라우드 기반의 서비스형 인공지능 (AI as a Service, AIaaS)

- **정의 :** 서비스형 인공지능(AIaaS)은 클라우드를 통해 즉시 사용할 수 있는 AI 서비스를 의미함
  - **경제성 :** 머신러닝 모델 개발을 위한 인프라, 엔지니어, 인건비, 데이터 등 비용을 절감할 수 있음
  - **접근성 :** 클라우드 내에서 공급업체가 관리하므로 언제, 어디서나 서비스에 접근 가능함 (AI 스피커)
  - **구현성 :** 사전에 개발된 인공지능 엔진을 제공하므로, 별도로 개발하거나 설치할 필요 없이 빠르게 적용할 수 있음

AIaaS가

사용자에게 제공할 수 있는 가치

싸게 (경제성)

쉽게 (접근성)

빠르게 (구현성)

Introduction

OKESTRO

❖ 클라우드 기반의 서비스형 인공지능 (AI as a Service, AIaaS)

- 서비스형 인공지능(AIaaS)은 **인공지능 기술을 직접 개발하지 않고도 애플리케이션에 인공지능 기술을 도입할 수 있음**
- 이러한 장점으로 **다양한 분야에서 지능형 애플리케이션 연구 개발에 도입**되고 있음

마케팅 도메인에서 서비스형 인공지능 활용의 분류의 예

Mechanical AI
Data Collection

- 시장, 기업, 경쟁사 및 고객에 대한 **데이터 수집을 자동화**
- **쉽게 자동화**될 수 있는 일상적이고 반복적인 작업

⇒ 데이터 감지, 추적, 수집

Thinking AI
Market Analysis

- 시장에 대한 **인사이트를 도출**
- 기존 시장 또는 새로운 시장을 **식별, 탐색, 예측**

⇒ 시장 이해 및 인사이트 도출

Feeling AI
Customer Understanding

- 기존 및 잠재적 고객의 **요구 이해**
- 기존 및 잠재적 고객에 **공감**

⇒ 고객 이해

Introduction

❖ 인공지능 기반 SWOT 분류 자동화

- 본 연구는 추후 클라우드 기반 서비스형 인공지능 플랫폼에서 상용화 될 수 있는 마케팅 인공지능 서비스 개발의 일환으로, 인공지능을 통해 **SWOT 분류를 자동화**하는 방식을 제안함

BI 관점에서의 문제 정의

Needs

직관적 기업 분석 자료에 대한 수요

As-Is

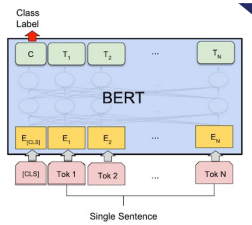
방대한 자료를 일일이 직접 읽고 정리

AI 기반 솔루션 연구

To-Be

실시간 자료 수집 자동화

실시간 크롤링으로 기업 데이터 수집



AI 데이터 분석

자체 수집 데이터로 학습한 AI로 자료 분석

AI 빅데이터 분석

SWOT, PEST, 3C, STP 등 다양한 비즈니스 전략 분석 자동화

Chapter 02

Approach

Approach

❖ 전체 아키텍처 (Overall Architecture)

- 본 연구는 SWOT 분류를 자동화하는 방식으로, 자연어추론(NLI) 방식으로 학습하는 **Sentence BERT**와 비지도학습 클러스터링 방식인 **K-Means 클러스터링** 알고리즘을 이용하여 SWOT 분류 문제를 풀고자 함

STEP - 01
자연어추론(NLI)

SWOT NLI 데이터 활용

SWOT Sentence BERT 학습

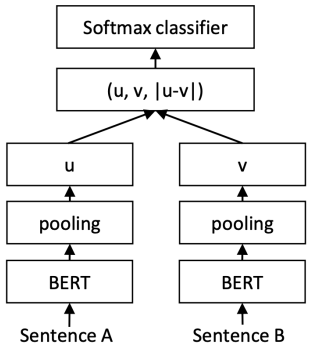


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

STEP - 02
특징 추출

모델

학습을 완료한
Sentence BERT

지도학습 데이터 활용

X = (S, W, O, T로 분류될 문장의 임베딩)

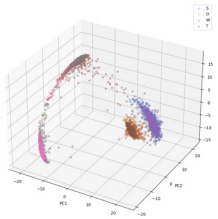
y = (S, W, O, T의 레이블)

STEP - 03
K-Means 클러스터링을 활용한 분류

학습 데이터

K-Means 학습(k=4)

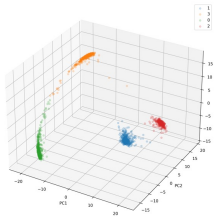
최빈 y값을 군집의 레이블로 가정



테스트 데이터

K-Means
클러스터 예측

가정한 군집의 레이블과 정답 y를 비교



Approach

❖ STEP 01 : 자연어추론(NLI)과 SWOT NLI dataset

- 자연어추론이란, 자연어이해(NLU)를 기반으로 모델의 추론 능력을 평가하는 작업으로, **언어 모델이 주어진 문장 간 의미를 파악하여 관계를 추론하는 문장 쌍 분류 문제**의 일종임

KLUE NLI

입력 Sentence 1 (전제)	입력 Sentence 2 (가설)	정답
24인치 캐리어 두개로 이미 여유공간은 없었다.	여유공간은 충분했다.	Contradiction (모순)
24인치 캐리어 두개로 이미 여유공간은 없었다.	그래서 캐리어를 세개 사용했다.	Neutral (중립)
24인치 캐리어 두개로 이미 여유공간은 없었다.	캐리어의 크기는 24인치였다.	Entailment (함의)

SWOT NLI

자체 수집한 약 1,358개 기업의 데이터 활용

입력 Sentence 1	입력 Sentence 2	정답
Strength 문장	Weakness 문장	Strength-Weakness
Weakness 문장	Threat 문장	Weakness-Threat
…	…	…
Opportunity 문장	Threat 문장	Opportunity-Threat

Approach

❖ STEP 02 : Sentence BERT(SBERT) 학습과 문장 특징 추출

- **Sentence BERT**란, 두 문장 간의 관계 학습을 통해 BERT의 문장 임베딩 성능을 우수하게 개선시킨 모델로, 파라미터를 공유하는 **Siamese 형태의 BERT**를 활용하여 **문장 쌍 분류(NLI)** 혹은 **문장 쌍 회귀(STS)** 태스크를 학습하는 모델임
- 학습을 마친 Sentence BERT는 **문장의 특징(Sentence Embeddings)을 추출하는 추출기**로써 활용됨

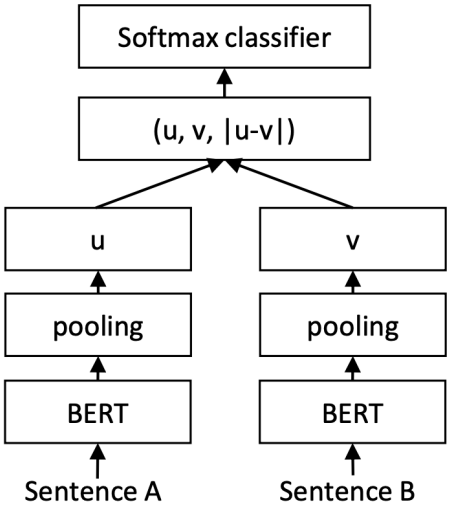


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

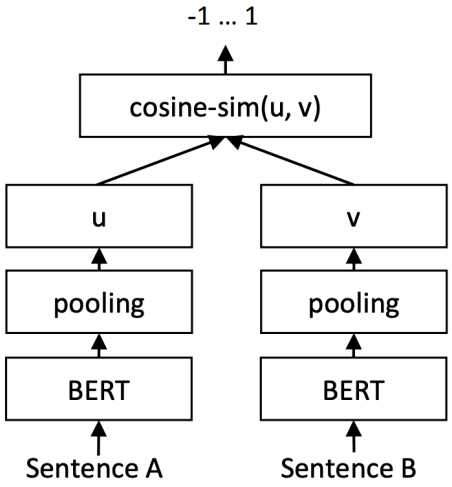


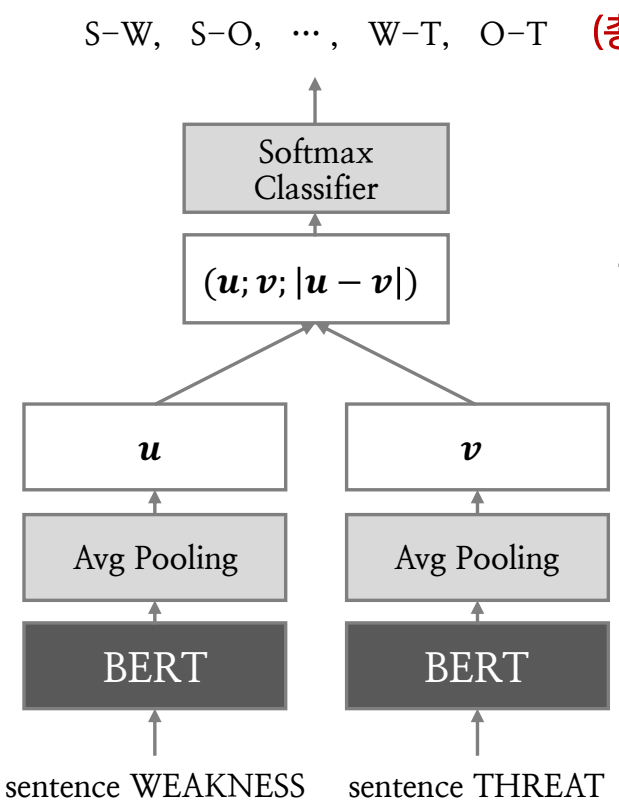
Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

Approach

❖ STEP 02 : Sentence BERT(SBERT) 학습과 문장 특징 추출

- **Sentence BERT**란, 두 문장 간의 관계 학습을 통해 BERT의 문장 임베딩 성능을 우수하게 개선시킨 모델로, 파라미터를 공유하는 **Siamese 형태의 BERT**를 활용하여 **문장 쌍 분류(NLI)** 혹은 **문장 쌍 회귀(STS)** 태스크를 학습하는 모델임
- 학습을 마친 Sentence BERT는 **문장의 특징(Sentence Embeddings)을 추출하는 추출기**로써 활용됨

SWOT NLI dataset으로 학습된
SWOT Sentence BERT



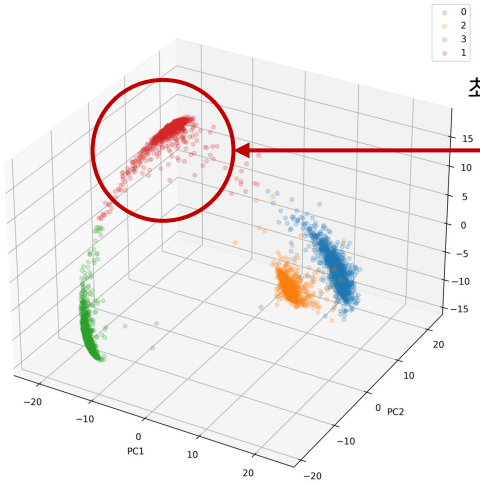
이후, SWOT Sentence BERT는
SWOT 문장의 특징 벡터를 추출하는
특징 추출기로 활용됨

Approach

❖ STEP 03-1 : K-Means Clustering과 정답 레이블 간주 (with training dataset)

- 분류 문제를 해결하기 위해, **지도 학습 형태로 SWOT dataset**을 재구축하였음
- 지도 학습 형태의 SWOT training dataset에 대해 **SWOT Sentence BERT를 통하여 문장 특징을 추출**함
- 추출된 문장 특징을 K-Means 클러스터링 하고, **해당 클러스터에 가장 많이 등장한 Ground Truth 레이블을 해당 클러스터에 대한 레이블로 가정**하였음
    - K-Means 클러스터의 개수는 각 문장이 S, W, O, T의 특징을 잘 표현할 것이라는 전제 하에 4개로 설정함

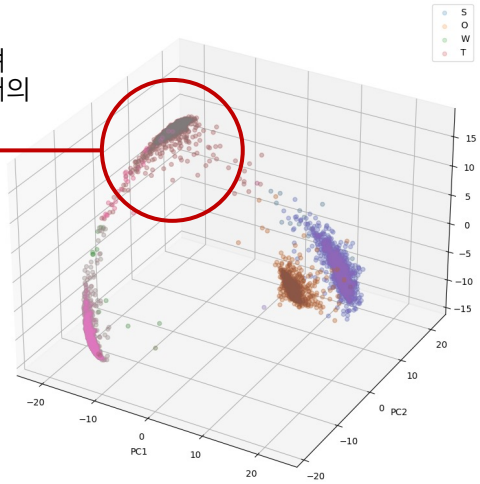
K-Means 클러스터 학습



지도 학습 training 데이터를 임베딩하여
K-Means 클러스터링(K=4) 학습

실제 정답과 비교하여
최빈 레이블을 클러스터의
정답으로 간주함

Ground Truth



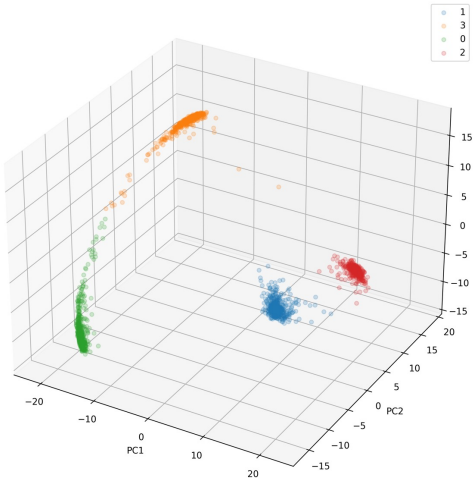
각 클러스터에서 가장 많이 등장한
실제 정답 레이블을 클러스터의 레이블로 간주함

Approach

❖ STEP 03-2 : K-Means Clustering을 활용한 분류 (with test dataset)

- 지도 학습 형태의 SWOT test dataset에 대해 **SWOT Sentence BERT를 통하여 문장 특징을 추출**함
- 추출된 문장 특징을 K-Means 클러스터를 예측하고, **해당 클러스터의 정답으로 간주했던 레이블을 예측 레이블로 가정**하였음
- 예측 레이블과 실제 test dataset의 레이블과 비교하였음

K-Means 클러스터 예측



지도 학습 test 데이터를 임베딩하여
K-Means 클러스터를 예측함

예측된 클러스터의 (간주)레이블과
실제 정답 레이블을 비교함

Chapter 03

Experimental Results

Experimental Results

OKESTRO

❖ 활용 데이터의 수가 늘수록 SBERT가 BERT를 능가함

- SWOT Sentence BERT와 K-Means 클러스터링의 분류 성능을 BERT의 지도학습 방식과 비교하였음
- 학습 데이터로 **활용한 기업의 수가 일정 수준(본 실험에서는 약 70개) 이상을 사용했을 때부터 SWOT Sentence BERT의 클래스 분류 F-1 점수가 BERT의 F-1점수보다 증가**하는 결과를 보여줌

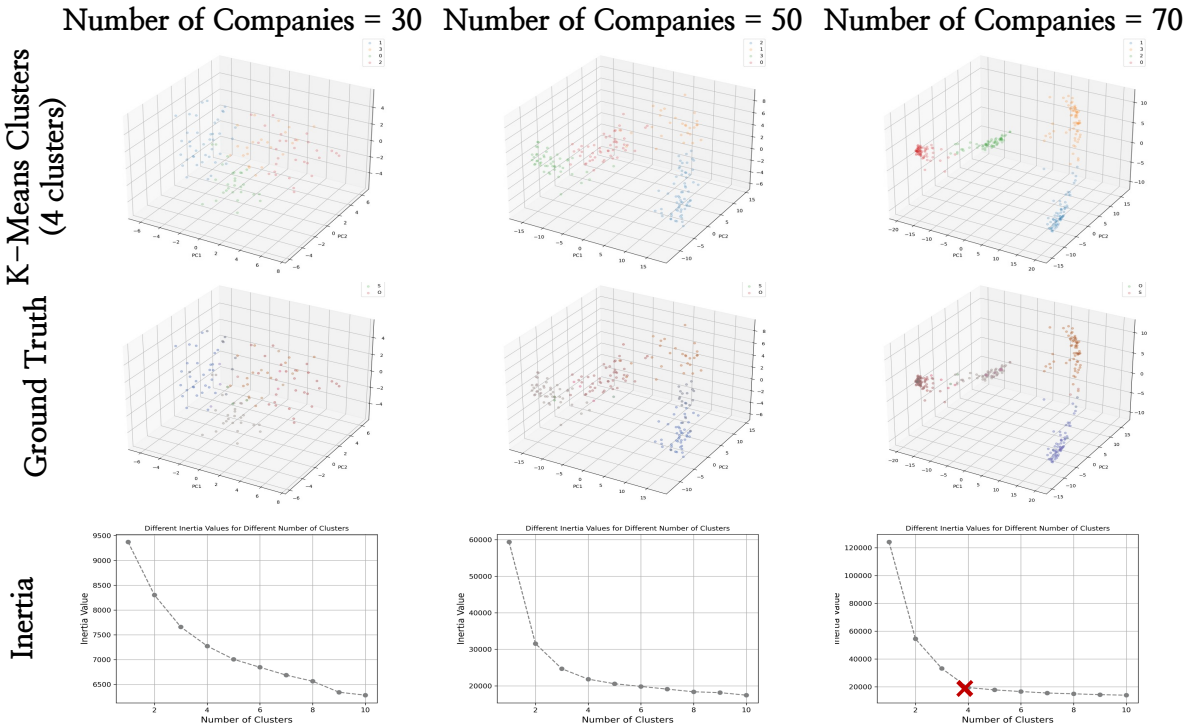
	Strength		Weakness		Opportunity		Threat		Weighted AVG	
# of Co.	BERT	SBERT	BERT	SBERT	BERT	SBERT	BERT	SBERT	BERT	SBERT
30	**0.8889**	0.75	0.6667	**0.7273**	0.8571	0.8571	0.7273	0.8	**0.7803**	0.7795
50	0.8421	**0.9412**	**0.9412**	0.8182	0.75	**0.8889**	**0.875**	0.6667	**0.8652**	0.8224
70	0.82	**0.93**	0.82	**0.9**	0.62	**0.83**	0.81	**0.97**	0.78	**0.93**
300	0.9524	**1**	0.907	**0.9545**	0.9048	**1**	0.88	**0.9615**	0.9093	**0.9774**
500	0.9536	**1**	0.8828	**0.9655**	0.8788	**1**	0.8421	**0.9682**	0.8897	**0.9827**
700	0.9321	**0.978**	0.8427	**0.9787**	0.8634	**0.9885**	0.8393	**0.9677**	0.8709	**0.9777**
1358	0.9261	**0.9944**	0.8815	**0.97**	0.8644	**0.9939**	0.8305	**0.971**	0.8745	**0.9813**

SWOT 분류 실험 결과 (F-1 점수)

Experimental Results

❖ 4개의 적정 클래스(S, W, O, T) 수로 수렴하는 클러스터의 관성

- 사용한 기업 데이터의 수가 일정 수준 이상이 되었을 때부터 클러스터의 적정 군집 수를 의미하는 **관성(Inertia)이 4로 결정**되는 것을 확인할 수 있음
- 이는 **SWOT Sentence BERT가 총 6개의 자연어추론 레이블로 학습되었음에도, 이를 통해 추출한 문장 임베딩의 클러스터가 정답 레이블의 특징을 충실히 반영**하는 것으로 해석할 수 있음



SWOT 데이터 K-Means(n_clusters=4) 클러스터(상),
실제 정답 데이터의 분포(중), 클러스터링 관성(Inertia)(하)

Chapter 04

Conclusions

Conclusions

❖ 본 논문의 의의

- 본 논문은 자연어 추론 방식으로 SWOT 데이터를 학습한 후 비지도학습 방식의 K-Means 클러스터링을 이용하는 **SWOT Sentence BERT가 일정 수준의 데이터가 확보될 경우 지도학습 방식의 BERT보다 SWOT 분류 자동화 성능이 우수함**을 보여줌

- 특히, **SWOT Sentence BERT가 최초 여섯 개의 클래스를 분류**하는 자연어 추론 방식으로 학습되었음에도 불구하고, 이를 통해 추출한 문장 임베딩의 클러스터는 정답 레이블인 **Strength, Weakness, Opportunity, Threat의 특징을 반영**하여 4개로 형성됨을 확인하였음

❖ 향후 활용 방안

- 추후 본 연구 내용을 기반으로 **클라우드 플랫폼을 통해 서비스형 인공지능으로 상용화**한다면, 마케팅 분석가의 수고가 들어가는 자연어 이해 기반의 SWOT 분류 작업을 자동화할 수 있는 애플리케이션으로 개발될 수 있을 것으로 기대됨

Conclusions

OKESTRO

❖ AI 기반의 SWOT 자동 분류 결과의 예시

Strength

"오케스트로는 클라우드 데이터센터를 구축하기 위한 **A부터 Z까지 풀스택 솔루션을 모두 보유**하고 있다. 김 총괄대표는 ""**글로벌 외산 제품도 구현하지 못한 기능들을 독자 기술력으로 구현**해 국내 뿐 아니라 많은 국제 **특허도 보유**하고 있다"며 기술력에 자신감을 보였다."

91%

Weakness

앞서 이 부회장은 2019년 "메모리에 이어 파운드리를 포함한 시스템 반도체 분야에서도 확실히 1등을 하겠다"는 비전을 제시한 바 있다. 당시 2030년까지 133조원을 투자하는 계획을 내놓았는데, 지난해 38조원을 더해 171조원으로 투자액을 늘렸다. 하지만 아직 제대로 된 성과는 내지 못했고, 오히려 격차가 벌어지는 상황이다. **삼성전자가 자랑해온 스마트폰 애플리케이션 프로세서(AP) '엑시노스 2200'은 올 초 낮은 성능과 수율 등의 문제가 발생**했고, 지난 4월 반도체 부문의 한 연구원이 이 부회장 등에게 전자우편을 보내 **경영진의 책임 회피와 패배의식 등 조직문화의 문제점**을 제기하기도 했다.

83%

기업 보도자료 기반, AI SWOT 분석 테스트 결과 샘플

Opportunity

"보고서를 작성한 강준영 교수는 ""반도체와 같이 대규모 투자와 연구개발에 오랜 시간이 소요되는 분야의 경우 정부가 인력·R&D·세제 등 전 분야에 걸쳐 연계하고 세밀하게 지원하는 게 필수적""이라고 말했다. 이어 ""**핵심 기술인력 확보의 경우 국내 우수인력 육성과 해외 핵심인력 유치를 동시에 진행**하고 있다""며 ""**한국이 정책 활용 차원에서 검토해볼 수 있을 것**""이라고 했다.

86%

Threat

지난해까지 **코로나19 팬데믹**으로 급증했던 **반도체 시장 수요가 올 들어 둔화 조짐**을 보이는 데다 **전 세계적인 고물가·고금리·고환율의 3고(高) 악재**가 겹치는 상황과 맞물려 **모바일과 PC 등 전방 수요가 둔화되면서 메모리 반도체 시장이 조정 국면에 들어갈 것**이라는 분석이다. 특히 서버수요도 **글로벌 경기 위축 여파**로 줄어들 가능성까지 제기된 상태다.

91%

References

❖ 참고 문헌

[1]Lins, Sebastian, et al. "Artificial Intelligence as a Service." Business & Information Systems Engineering 63.4 (2021): 441-456.

[2]Huang, Ming-Hui, and Roland T. Rust. "A strategic framework for artificial intelligence in marketing." Journal of the Academy of Marketing Science 49.1 (2021): 30-50.

[3]Reimers, Nils, and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks." arXiv preprint arXiv:1908.10084 (2019).

[4]Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[5]Park, Sungjoon, et al. "Klue: Korean language understanding evaluation." arXiv preprint arXiv:2105.09680 (2021).