# BERT 를 활용한 한국어 지속가능경영 보고서의 제로샷 가독성 평가

손규진 [1], 윤나은 [2], 이가은 [3*]
[1] 연세대학교 언더우드국제대학 경제학과
[2] 연세대학교 산업공학과
[3] 한림대학교 중국학과
spthsrbwls123@yonsei.ac.kr, na803704@yonsei.ac.kr, lgeun726@gmail.com

# Zero-Shot Readability Assessment of Korean ESG Reports using BERT

Guijin Son[1], Naeun Yoon[2], KaeunLee[3]
[1]Dept. of Economics, Underwood International College, Yonsei University
[2]Dept. of Industrial Engineering, Yonsei University,
[3]Dept. of Chinese, Hallym University

**요 약**

본 연구는 최근 자연어 인공지능 연구 동향에 발맞추어 사전 학습된 언어 인공지능을 활용한 의미론적 분석을 통해 국문 보고서의 가독성을 평가하는 방법론 두 가지를 제안한다. 연구진은 연구 과정에서 사전 학습된 언어 인공지능을 활용해 추가 학습 없이 문장을 임의의 벡터값으로 임베딩하고 이를 통해 *1. 의미론적 복잡도* 와 *2. 내재적 감정 변동성* 두 가지 지표를 추출한다. 나아가, 앞서 발견한 두 지표가 국문 보고서의 가독성과 정(+)의 상관관계에 있음을 확인하였다. 본 연구는 통사론적 분석과 레이블링 된 데이터에 크게 의존하던 기존의 가독성 평가 방법론으로부터 탈피해, 별도의 학습 없이 기존 가독성 지표에 근사한다는 점에서 의미가 있다.

## 1. 서 론

회계와 재무 분야 기업 공시 보고서의 가독성을 평가하고 대중에게 쉽게 읽히는 공시 자료를 만들기 위한 노력은 산업과 학계 양측에서 지속적으로 이루어져 왔다. 안타깝게도 그간의 연구는 대개 기술적 한계로 인해 음절의 길이나 특정 품사의 수와 같이 통사론적 특징을 매개 변수로 가독성을 평가해 왔으며, 의미론적 분석을 통한 가독성 평가에 관한 연구는 미비한 상황이다 [1].

그러나, 2018년 대량의 자연어 말뭉치에 사전 학습된 언어 인공지능 BERT의 등장 이후로 인공지능을 활용해 언어의 의미론적(semantic) 특징을 벡터값으로 인코딩하는 것이 가능해졌다 [2]. 본 연구는 사전 학습된 언어 인공지능이 출력하는 벡터값을 추가 학습 없이 활용하여 한국어 지속가능경영 보고서의 가독성을 평가하는 제로샷 방법론 두 가지를 소개한다[2]. 본 연구 결과는 통사론적 분석과 레이블링 되어 있는 데이터에 크게 의존하던 기존의 가독성 평가 연구와 달리 추가적인 학습 없이 가독성을 평가한다는 점에서 의미를 가진다.

## 2. 선행 연구

2 장에서는 가독성 평가를 위해 기존에 가장 널리 사용되던 지표와 사전 학습된 언어 인공지능 BERT에 관한 선행 연구를 소개한다.

### 2.1 가독성 평가 지표에 관한 선행 연구

가장 널리 사용되는 통사론적 가독성 평가 방법론으로는 Gunning Fog Index (FOG)와 Flesch Reading Ease Formula (FLESCH)가 존재한다. FOG 는 문장의 평균 길이(단어 수/문장 수)와 3 음절

* 현재 졸업 이후 무소속
[2] 연구에 사용한 모든 코드는 아래에 첨부함:
https://github.com/guijinSON/Korean_Readability_Assessment

이상 단어 수의 합으로 계산되며, 산출식은 아래와 같다 [3].

$$FOG = (\textit{문장의 평균 길이} + 3\ \textit{음절 이상 단어 수}) * 0.4$$

<수식 1> Gunning Fog Index 산출식

영어와 음절의 구조가 다른 한글을 대상으로 하는 연구의 경우 5 음절, 7 음절, 10 음절 이상 단어 수로 FOG 를 측정하여 사용하기도 한다 [1].

이와 유사하게 FLESCH 는 단어의 평균 길이(총 음절 수/ 총 단어 수)와 문장의 평균 길이에 기반하여 계산되며, 산출식은 아래와 같다 [4].

$$FLESCH = 206.835 - (84.6 * \textit{단어의평균길이}) - (1.015 * \textit{문장의평균길이})$$

<수식 2> Flesch Reading Ease Formula 산출식

본 논문에서는 다른 가독성 지표와 같은 방향으로 해석하기 위해 FLESCH 에 -1 을 곱하여 측정하였다. 이때, FOG 와 FLESCH 모두 값이 커질수록 보고서의 가독성이 낮아짐을 의미한다.

### 2.2 사전 학습된 언어 인공지능에 관한 선행 연구

사전 학습된 언어 인공지능은 많게는 테라바이트(TB)에 이르는 대용량의 말뭉치 데이터를 사용해 언어에 대한 범용적 지식을 습득한 인공지능을 의미한다. 빈칸 단어 예측(Masked Language Modeling)과 다음 문장 예측(Next Sentence Prediction) 등 데이터 레이블링을 필요로 하지 않는 자기지도(self-supervised) 과제를 통해 언어의 의미론적(semantic), 통사론적(syntactic) 특징을 학습한다 [2]. 사전 학습된 언어 인공지능은 가독성 평가를 위해서도 활발히 활용되고 있다. 그러나, 대부분 지문과 사람이 직접 평가한 가독성 점수가 레이블링 된 데이터에 추가 학습시켜 진행하며 이와 같은 경우 레이블링 된 데이터가 필수적이라는 한계가 존재한다 [5]. 이를 보완하기 위해 사전 학습된 언어 인공지능과 자기 회귀모델을 사용해 추가 학습 없이 가독성을 평가할 수 있는 Ranked Sentence Readability Score (RSRS)가 최근 활발히 연구되고 있다 [6].

RSRS 는 사전 학습된 언어 모델이 학습한 언어의 분포가 지프의 법칙을 따라, 표준에 가깝고 가독성이 높은 문장은 낮은 펄플렉시티 (perplexity)를 가지고, 복잡하고 희귀한 언어 구조를 가진 문장의 펄플렉시티가 높다는 점에서 기인한다. RSRS 는 문장의 각 단어 별로 음의 로그우도 값(negative log-likelihood score)[3] 을 기반으로 계산된다. 문장 내 단어 수만큼 계산된 음의 로그우도 값들을 오름차순으로 정렬하여 순위를 매기고, 순위를 제곱근 한 값을 곱해준다. 이는 가독성에 영향을 많이 미치는 단어에 가중치를 주기 위함이다. 개별 문장의 RSRS 산출식은 아래와 같다.

$$RSRS = \frac{\sum_{i=1}^{\textit{문장의 길이}} \sqrt{i} * (i\ \textit{번째 음의 로그우도 값})}{\textit{문장의 길이}}$$

<수식 3> Ranked Sentence Readability Score 산출식

위 공식을 사용해 개별 문장의 가독성을 계산하고 이를 평균 내어 문서 전체의 가독성을 산출한다는 점에서 RSRS 는 FOG 와 유사점을 가진다.

## 3. 연구 설계

### 3.1 데이터셋 구축 과정

본 연구는 한국표준협회(KSA) [4] 에서 수집한 지속가능성 보고서와 연합뉴스에서 수집한 경제 기사를 활용한다 [7]. 크롤링을 통해 수집한 텍스트 데이터는 아래의 전처리 과정을 거쳐 가독성을 평가하기 적합한 형태로 변형되었다.

1. (종결어미+마침표)로 마무리되는 문장만 추출.
2. 문장 띄어쓰기 교정.
3. 숫자, 특수문자, 외국어 제거.

KSA 에서 수집한 지속가능성 보고서는 PDF 형태의 파일로 Tika[5] 라이브러리를 통해 텍스트로 변환했다. 이후, 제목, 캡션, 표 등을 제거하기 위해 (종결어미 + 마침표)로 마무리되는 문장만을 추출하였다. 여타 자연어 처리 과제와 달리 가독성 평가의 경우 문장 및 단어의 길이에 크게 의존하므로 앞서 추출한 문장의 띄어쓰기를 Pororo [6] 라이브러리로 재차 교정하였다. 마지막으로 고유명사를 표현하기 위해 사용된 외국어는 가독성을 저해하지 않는다고 판단해 모두 제거해 주었다. 전처리 과정이 완료된 이후 데이터셋에 대한 정보는 아래 <표 I> 과 같다.

| 데이터 종류 | 총 개수 | 평균 길이 |
|---|---|---|
| 지속가능성 보고서 | 1,387 개 | 16,879 자 |
| 경제 기사 | 1,500 개 | 917 자 |

<표 I> 데이터셋 상세 정보

[3] 퍼플렉시티=Exponential(음의 로그우도 값)
[4] 한국표준협회: https://www.ksa.or.kr/ksa_kr/index.do
[5] Tika 라이브러리: https://tika.apache.org
[6] PORORO: Platform Of neuRal mOdels for natuRal language prOcessing: https://github.com/kakaobrain/pororo

| | FOG (5 음절) | FOG (7 음절) | FOG (10 음절) | RSRS | 의미론적 복잡도 (biRSRS) | P-Value |
|---|---|---|---|---|---|---|
| 문서 길이 | 0.094883 | 0.088619 | 0.086788 | 0.089638 | **0.028542** | 0.688 |
| FOG(5 음절) | 1 | 0.995612 | 0.995429 | 0.675629 | **0.747636** | $10^{-10}$ |
| FOG(7 음절) | - | 1 | 0.99986 | 0.630768 | **0.726106** | $10^{-10}$ |
| FOG(10 음절) | - | - | 1 | 0.629978 | **0.724619** | $10^{-10}$ |
| RSRS | - | - | - | 1 | **0.732547** | $10^{-10}$ |
| 의미론적 복잡도 (biRSRS) | - | - | - | - | 1 | - |

<표 II> 가독성 평가 지표 간의 상관계수

### 3.2 의미론적 복잡도

의미론적 복잡도 (bidirectional-RSRS, biRSRS)는 본 연구진이 기존의 RSRS 지표가 양방향성을 가지도록 개선한 것으로 우도 계산이 이전보다 더 많은 정보에 기반케 한다. 기존의 RSRS 는 자기 회귀 구조를 사용해 우도를 계산하고자 하는 목표 단어 이전에 등장한 문맥만을 활용한다. 그러나, BERT 모델을 사용하는 biRSRS 의 경우 목표 단어가 속해 있는 문장 전체를 기반으로 우도를 산출한다. [그림 1]은 자기 회귀 모델을 활용한 RSRS 의 작동 방식을, [그림 2]는 양방향성을 가지는 biRSRS 의 작동 방식을 설명한다.

[그림 1] RSRS 의 작동방식

[그림 2] biRSRS 의 작동방식

biRSRS 와 RSRS 는 FOG 와 동일하게 개별 문장의 가독성을 평가해 평균을 내는 방식을 사용한다. 위 특성을 고려하여 본 연구에서는 RSRS 와 비교했을 때 FOG 와 더 높은 상관관계를 가지는 개선된 biRSRS 를 만드는 것을 목표한다. 이때, 본 연구에서는 문장의 biRSRS 를 생성하기 위해 Kakao 가 학습한 brainbert 모델을 사용하였다.

### 3.3 내재적 감정 변동성

문서의 내재적 감정 변동성 (Document Sentiment Volatility)은 본 논문의 연구진이 아는 한 감정 분석(Sentiment Analysis)을 가독성 평가에 활용하는 최초의 시도이다. 내재적 감정 변동성은 일관된 감정을 주장하는 보고서에 비하여 연속된 문장 사이에 감정 변화가 클 경우 가독성을 해 할 것이라는 가설에 기반하여 제작되었다. 내재적 감정 변동성의 산출 방식은 아래와 같다.

1. 문서를 문장 단위로 분리.
2. 각 문장의 감정 점수 (sentiment score) 계산.
3. 감정 변동성 산출.

각 문장의 감정 점수는 NSMC[7] 데이터셋을 통해 긍부정 분류 과제를 추가 학습한 BERT 모델을 사용하여 계산되었다. 이후, 개별 문장 단위로 계산된 감정 점수를 하나의 감정 시계열로 만들어 변동성을 산출하며, 이를 위해 아래 공식을 사용한다.

$$\textit{내재적 감정 변동성} = \textit{감정 시계열의 표준편차} * \sqrt{\textit{감정 시계열의 길이}}$$

<수식 4> 내재적 감정 변동성 산출식

본 논문이 새로이 제안하는 내재적 감정 변동성 지표는 단순히 개별 문장 단위의 가독성 점수를 평균낸 FOG 나 RSRS 와는 달리 연결된 문장 간의 관계에 기반해 가독성을 평가한다는 점에서 차이를 가진다. 이 때문에, 내재적 감정 변동성은 FOG 와의 상관 계수 대신 금융 기사와 지속가능경영 보고서의 가독성을 각각 계산하고 그 결과를 바탕으로 성능을 평가한다.

## 4. 연구 결과

4 장에서는 본 논문에서 제안하는 두 지표 의미론적 복잡도와 내재적 감정 변동성의 지속가능경영 보고서 가독성 평가 결과를 확인한다. 모든 성능 평가는 결과의 공평함을 위해 동일한 Tesla P-100 하드웨어 가속기를 활용해 진행되었다.

[7] Naver Sentiment Movie Corpus: https://github.com/e9t/nsmc

#### 4.1 의미론적 복잡도

본 연구가 제안하는 의미론적 복잡도(biRSRS)는 FOG 와 동일한 방식으로 가독성을 계산하며, 이에 FOG 와의 상관성을 통해 성능을 평가한다. <표 II> 는 문서의 길이, FOG(5,7,10 어절), RSRS, biRSRS, 내재적 감정 변동성 총 7 가지의 변수 사이의 상관성을 피어슨 상관 계수로 표현한 표이다. 분석 결과 개선 이전의 RSRS 와 비교했을 때 개선된 biRSRS 는 FOG 와 1 에 더 가까운 양의 선형 상관관계에 있음을 확인할 수 있다. <표 II>의 마지막 열은 biRSRS 지표의 상관계수가 가지는 통계적 유의미함을 P-Value 로 측정한 것으로 통상적으로 P-Value 가 0.01 이하 일때 통계적으로 유의미한 상관이 있다고 간주한다. 본 실험에서는 biRSRS 와 기존의 FOG(5,7,10 어절)는 모두 약 $10^{-10}$ 이하의 P-Value 값을 보였다.

#### 4.2 내재적 감정 변동성

FOG 와의 상관관계를 통해 성능을 분석한 의미론적 복잡동와는 달리 내재적 감정 변동성은 문장과 문장 사이의 변화에 기반해 가독성을 측정하며, 그 때문에 FOG 와 상관성이 부족한 것은 당연한 현상이라고 생각된다. 이에 본 연구진은 내재적 감정 변동성의 성능을 평가하기 위해 앞서 수집한 지속가능경영 보고서와 금융 기사를 활용한다.

| | FOG (5 음절) | FOG (7 음절) | 내재적 감정 변동성 |
|---|---|---|---|
| 지속가능경영보고서 | 12.86 | 12.26 | 3.31 |
| 금융 기사 | 4.90 | 4.56 | 1.12 |
| *지속가능경영 보고서 / 금융 기사* | **2.63** | **2.69** | **2.96** |

<표 III> 내재적 감정 변동성 가독성 평가

<표 III> 은 FOG(5,7 음절) 과 내재적 감정 변동성으로 앞서 수집한 지속가능경영 보고서와 금융 기사 각각의 가독성을 평가한 것이다. 비록 두 지표가 서로 다른 척도(scale)을 사용하지만 *(지속가능경영 보고서/금융 기사)*의 값을 보면 세 지표 모두 동일하게 금융 기사의 가독성이 지속가능경영 보고서보다 약 3 배 정도 나음을 보이고 있다. 이를 바탕으로 본 연구진은 과정만 다를 뿐 내재적 감정 변동성 역시 기존의 가독성 지표와 유사한 결론에 다다를 수 있음을 확인하였다.

## 5. 결론

본 논문에서는 사전 학습된 언어 인공지능 BERT 를 사용하여 별도의 학습 없이 기존의 가독성 지표에 근사하는 두 종의 새로운 지표, 의미론적 복잡도와 내재적 감정 변동성을 소개한다. 실험 결과 두 지표 모두 레이블링 된 데이터를 활용하는 추가 학습 없이 기존의 가독성 지표에 근사하는 모습을 보여주었다. 이와 같이 두 지표 모두 사전 학습된 모델을 추가 비용 없이 사용하여 결과를 내는 "제로샷 모델"로서의 특징을 가진다는 점에서 의미가 있다.

금융 분야의 디지털 전환 트렌드와 함께 회계와 재무 분야 기업 공시 보고서의 가독성을 인공지능으로 평가하고 대중에게 쉽게 읽히는 공시 자료를 만들기 위한 노력은 꾸준히 이루어지고 있다. 그러나 대량의 데이터를 레이블링하고 이를 바탕으로 초대형 모델을 재학습하는 기존의 방식에는 큰 비용이 발생한다. 따라서, 본 연구와 같이 언어 인공지능이 사전 학습 단계에서 습득한 지식을 최대한으로 활용하여 위 비용을 절약하고자 하는 시도는 계속되어야 한다.

## 참고문헌

[1] 정태진, 임승연, 이우종, 조미옥. "우리말 사업보고서 가독성 연구의 가능성에 대한탐색적 연구." 회계학연구 43.4 pp.37-100. 2018.

[2] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) pp. 4171-4186. 2019

[3] Gunning Robert. "The Technique of Clear Writing." New York. New McGraw-Hill Book Co. 1952.

[4] Flesch, R. "A New Readability Yardstick." Journal of Applied Psychology 32 (3): 221-233. 1948.

[5] Filighera, Anna, Tim Steuer, and Christoph Rensing. "Automatic text difficulty estimation using embeddings and neural networks."European Conference on Technology Enhanced Learning. Springer, Cham, 2019.

[6] Martinc, M., Pollak, S., & Robnik-Šikonja, M. "Supervised and unsupervised neural approaches to text readability." *Computational Linguistics, 47 (1), 141-179. 2021.*

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