JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 24 No. 11, pp. 41-49, November 2019
https://doi.org/10.9708/jksci.2019.24.11.041

The Unsupervised Learning-based Language Modeling of Word Comprehension in Korean

Euhee Kim*

*Professor, Dept. of Computer Science & Engineering, Shinhan University, Gyeonggi, Korea

[Abstract]

We are to build an unsupervised machine learning-based language model which can estimate the amount of information that are in need to process words consisting of subword-level morphemes and syllables. We are then to investigate whether the reading times of words reflecting their morphemic and syllabic structures are predicted by an information-theoretic measure such as surprisal. Specifically, the proposed Morfessor-based unsupervised machine learning model is first to be trained on the large dataset of sentences on Sejong Corpus and is then to be applied to estimate the information-theoretic measure on each word in the test data of Korean words. The reading times of the words in the test data are to be recruited from Korean Lexicon Project (KLP) Database. A comparison between the information-theoretic measures of the words in point and the corresponding reading times by using a linear mixed effect model reveals a reliable correlation between surprisal and reading time. We conclude that surprisal is positively related to the processing effort (i.e. reading time), confirming the surprisal hypothesis.

▸ **Key words**: Unsupervised learning, Morfessor, Surprisal, Lexical processing, Word recognition

[요 약]

본 연구는 비지도 기계학습 기술과 코퍼스의 각 단어를 이용하여 한국어 단어를 형태소 분석하는 언어 모델을 구축하는데 목적을 둔다. 그리고 이 언어 모델의 단어 형태소 분석의 결과와 언어 심리 실험결과에서 얻은 한국어 언어사용자의 단어 이해/판단 시간이 상관관계을 갖는지를 규명하고자 한다. 논문에서는 한국어 세종코퍼스를 언어 모델로 학습하여 형태소 분리 규칙을 통해 한국어 단어를 자동 분리하는데 발생하는 단어 정보량(즉, surprisal(놀라움) 정도)을 측정하여 실제 단어를 읽는데 걸리는 반응 시간과 상관이 있는지 분석하였다. 이를 위해 코퍼스에서 단어에 대한 형태 구조 정보를 파악하기 위해 Morfessor 알고리즘을 적용하여 단어의 하위 단위 분리와 관련한 문법/패턴을 추출하고 형태소를 분석하는 언어 모델이 예측하는 정보량과 반응 시간 사이의 상관관계를 알아보기 위하여 선형 혼합 회귀(linear mixed regression) 모형을 설계하였다. 제안된 비지도 기계학습의 언어 모델은 파생단어를 d-형태소로 분석해서 파생단어의 음절의 형태로 처리를 하였다. 파생단어를 처리하는 데 필요한 사람의 인지 노력의 양 즉, 판독 시간 효과가 실제로 형태소 분류하는 기계학습 모델에 의한 단어 처리/이해로부터 초래될 수 있는 놀라움과 상관함을 보여 주었다. 본 연구는 놀라움의 가설 즉, 놀라움 효과는 단어 읽기 또는 처리 인지 노력과 관련이 있다는 가설을 뒷받침함을 확인하였다.

▸ **주제어**: 비지도 학습, 모페써, 놀라움, 선형 혼합 회귀 모형, 단어 판독

• First Author: Euhee Kim, Corresponding Author: Euhee Kim
*Euhee Kim (euhkim@shinhan.ac.kr), Dept. of Computer Science & Engineering, Shinhan University
• Received: 2019. 10. 07, Revised: 2019. 11. 07, Accepted: 2019. 11. 07.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

본 연구는 비지도 기계학습 기술과 코퍼스의 각 단어를 이용하여 한국어 단어를 형태소 분석하는 언어 모델을 구축하는데 목적을 둔다. 그리고 이 언어 모델의 단어 형태소 분석의 결과와 언어 심리 실험 결과에서 얻은 한국어 언어사용자의 단어 이해/판단 시간이 상관관계를 갖는지를 규명하고자 한다.

자연언어 처리 응용 분야에서 코퍼스 기반 연구는 컴퓨터 등장 이후 보편화된 것으로 최근 빅데이터 연구 및 데이터 사이언스에서도 코퍼스를 수집하여 분석하고 있다. 본 연구와 관련한 기존 연구로서 트위터나 페이스북 같은 SNS에 등록한 사용자 계정 이름을 수집하여 비지도 학습을 통해 언어 모델을 구축하였다. 이 모델을 이용하여 영어사용자 계정 이름만을 사용해서 사용자 성별 및 연령대 관계를 추론하였다. 예를 들어, 트위터 계정 이름이 @taylorswift13일 경우 “tayler”, “swift”, “13”의 하위 단위로 분리되는 경우 각 하위 단위에는 개인의 정보 관련 의미를 함의할 수 있다. 이런 분류를 자동화하기 위해 Cruetz와 Lagus가 제안한 Morfessor 알고리즘을 사용하여 사용 계정 이름을 비지도 학습을 통해 하위 단위로 분리하였다. 또한, 모델의 성능을 검증하기 위해 n-gram 기본 모델과 비교하였다[1-2].

언어 인지과학에서는 단어를 처리할 때 시각적 방법을 사용하여 단어 인식 과제에서 반응 시간 측정, 단어 읽기 중 안구 운동의 추적, 단어의 시각적 표현으로 도출된 뇌 활동을 측정하는 기술과 같은 다양한 도구를 활용하여 연구되어지고 있다. 관련 최근 연구로서, 영어 단어를 읽을 때 안구 운동을 추적하는 툴을 사용하여 어휘 처리를 분석한 실험 결과와 Morfessor 알고리즘을 사용하여 단어를 하위 단위로 분리하여 어휘 판단 처리를 예측할 수 있는 코퍼스를 이용한 통계 기반 언어 모델의 처리 결과를 비교하였다[3-4].

그러나 아직 한국어 단어 코퍼스를 대상으로 비지도 기계학습 모델을 이용하여 형태적으로 복잡한 파생단어 인식을 예측하는 연구는 현재까지 찾아보기가 쉽지 않다. 특히 자연어 처리 관련 연구가 활발한 영어와는 달리, 한국어의 경우 상응하는 관련 자료가 적기 때문에 이와 같은 연구가 부족한 실정이다.

위에서 언급한 기존 연구를 바탕으로 본 연구에서는 한국어 코퍼스를 언어 모델로 학습하여 형태소 분리 규칙을 통해 한국어 단어를 자동 분리하는데 발생하는 단어 정보량(즉, surprisal(놀라움) 정도)을 측정하여 실제 단어를 읽는데 걸리는 반응 시간과 상관이 있는지 분석하고자 한다. 이를 위해 코퍼스에서 단어에 대한 형태 구조 정보를 파악하기 위해 Morfessor 알고리즘을 적용하여 단어의 하위 단위 분리와 관련한 문법/패턴을 추출하고 형태소를 분석하는 언어 모델이 예측하는 정보량과 반응 시간 사이의 상관관계를 알아보기 위하여 선형 혼합 회귀(linear mixed regression) 모형을 설계한다.

본 논문의 구성은 다음과 같다. 2장에서는 형태소 분석 언어 모델에 사용할 Morfessor 알고리즘과 선형 혼합 회귀 모형을 기술한다. 3장에서는 언어 실험 대상이 되는 데이터와 코퍼스 기반 형태소 분석 언어 모델을 제시하며, 4장에서는 제시한 모델의 실험 결과를 분석한다. 5장은 실험 결과에 대한 요약 및 향후 과제를 제시한다.

II. Methods

1. d-morph(ology)

형태소는 언어학에서 특정한 어휘나 문법 의미를 가지는 가장 작은 말의 단위로 따로 분리할 수 있는 것을 말한다. 단어의 형태 구조를 반영하는 형태소 분석기는 단어를 보고 형태소 단위로 분리해내는 소프트웨어를 말한다. 이러한 형태소 분석은 자연언어 처리의 가장 기초적인 절차로 구문 분석이나 의미 분석을 수행하기 전의 언어 분석 과정이다. 기존의 형태소 분석기는 규칙 기반으로 동작하기 때문에 사람이 직접 지속적으로 규칙을 입력해야 한다. 세종코퍼스의 Kmma 형태소 분석기나 Utagger 형태소 분석기는 문장을 입력하면 품사를 기반으로 문장을 구성하는 단어들에 대한 형태소 분석 결과를 얻을 수 있다. 또한, 최근에 데이터 기반으로 동작하는 딥러닝 알고리즘을 사용하여 카카오에서 개발한 khaiii 형태소 분석기가 있다[5-9].

형태소들을 다양한 방법으로 결합하여 형태적으로 복잡한 단어 혹은 파생단어를 생성한다. 예를 들어, 문법 의미를 고려하지 않고, 단어 “가계부”를 잠재적으로 “가계”+“부”, “가”+“계부”, 그리고 “가”+“계”+“부”로 분리할 수 있다. 본 연구에서는 언어학에서 사용하는 형태소 개념과 구분하여 사용하기 위해 파생단어를 구성하는 음절들을 최소 단위로 사용하여 분리하는 방법을 데이터 기반 d-형태소(d-morph(ology))라 정의한다.

본 연구를 위해 파생단어에 대한 d-형태소 분석은 기존의 형태소 분석기를 사용할 수가 없다. 따라서 파생단어를 d-형태소로 자동 분리하기 위해서 한국어 코퍼스를 대상으로 형태소 분석하는 문법/패턴을 추출하기 위해 Morfessor 알고리즘을 사용하였다.

2. Unsupervised d-Morphology Learning

기존 형태소 분석기 버전이 사전과 규칙에 기반을 두어 분석을 하는데 비해, Morfessor 알고리즘은 코퍼스를 대상으로 학습하여 d-형태소 분석을 수행한다. 학습에 사용한 코퍼스는 국립국어원에서 배포한 21세기 세종코퍼스를 선택하였으며, 이 코퍼스의 오류를 수정하여 사용하였다.

Creutz와 Lagus가 제안한 Morfessor 알고리즘은 원시 텍스트 데이터만을 이용하여 음절 기반으로 형태소를 분리하는 비지도 확률 기계학습 알고리즘이다. 이 알고리즘은 최적의 형태소를 찾기 위해 반복 기법을 사용한다.

Morfessor 알고리즘은 음성 인식이나 기계 번역 등 자연어 처리 응용 분야에서 비영리적으로 사용이 가능하다. 영어와 달리 한국어는 교착어에 속한다. 즉, 어근과 접사에 의해 단어의 기능이 결정되는 언어의 형태이다. 따라서 한국어의 경우 고유 단어의 수가 기하급수적으로 증가하게 된다. 이런 한계를 극복하기 위해 Morfessor 알고리즘을 사용하여 한국어 교착어를 입력하면, 교착어로 이루어진 텍스트 안에서 음절을 기반으로 형태소를 찾아내고, 최소의 형태소로 단어를 어떻게 분리하여 하는지를 비지도 학습을 통해 자동으로 찾아낸다.

Morfessor 2.0 알고리즘을 이용한 데이터 기반 형태소 분석 모델의 동작 원리는 그림 1과 같이 학습 파생단어 코퍼스를 대상으로 최적의 d-형태소(a lexicon of d-morphs)와 형태소 분리 규칙(constructions)을 생성하는 과정을 학습단계와 테스트 단계로 나뉜다. 여기서, derived words는 파생단어을 의미한다. 형태소 분리 규칙과 파생단어의 가장 작은 형태는 문자가 된다. 모델을 검증하기 위해서, d-형태소로 태깅된 데이터(annotated training data)를 무작위로 1000 개의 샘플을 추출하여 Morfessor 알고리즘을 훈련할 때 준지도 검증 데이터로 사용하였다. 모델을 최적화시키는 Minimum Description Length(MDL) 목적함수를 갖고 있다.

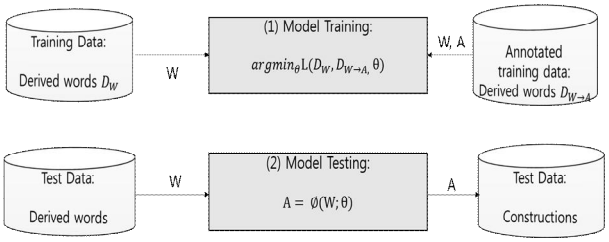

Fig. 1. The Workflow for Morfessor

이 연구의 목적을 위해 한국어 어휘를 d-형태소로 변환하는 Morfessor 2.0 모듈로 사용했다. 그림 1에서 기술한 것처럼, 한국어 파생단어에 대한 최적의 d-형태소로 분리된 n-gram에 대한 확률 $\phi(W;\theta)$은 파생단어 W가 d-형태소 $m_1, m_2, ..., m_n$으로 분리된 다음 비용함수를 최소화시키는 파라미터 θ를 찾는다. Morfessor기반 언어 모델을 구축하기 위해 2-grams과 3-grams 자질을 사용하였다. 본 논문에서는 Python의 Morfessor 2.0 패키지를 사용하여 구현하였다[10-11].

3. Word-Information metric

파생단어는 k개 d-형태소로 구성되었다고 가정할 때, 예를 들어, 3개 d-형태소로 구성된 파생단어는 $w = m_1 m_2 m_3$로 표현 할 수 있다. 효율적인 언어 모델은 실제 사람이 사용하는 언어와 최대한 비슷하게 확률 분포를 근사하는 모델이다. 많이 사용하는 단어일수록 확률을 높게 예측해야 하며, 적게 사용하는 단어는 확률을 낮게 예측해야 한다. d-형태소의 개수가 k인 파생단어 $(m_1, m_2, \cdots, m_k)$에 대한 예측 과정은 단어의 앞부분이 주어지고, 다음에 나타나는 음절의 확률 분포가 실제 테스트 단어의 다음 음절에 대해 높은 확률을 갖는다면 성능이 좋은 언어모델이라 할 수 있다.

단어의 형태소 분석에 대한 놀라움은 임의 음절에 대한 불확실성을 측정한 것으로 t번째 음절에 대한 정보량을 이용하여 실제 다음 t+1 번째 음절의 확률로 정량화한다. 음절의 놀라움은 그 음절의 예상치 못한 정도를 측정한다고 말할 수 있다. 단어의 형태소 분석에 대한 정보량을 놀라움의 정도를 계산하기 위해 그림 1의 $\phi(W;\theta)$을 사용하여 다음과 같이 계산되어진다.

$$surprisal(m_{t+1}) = -log_e \emptyset_{best}(m_{t+1}|m_1, \ldots, m_t)$$

Fig. 2. Surprisal for a Derived Word

본 연구에서는 파생단어에 대한 Morfessor 알고리즘의 log-probability을 이용하여 단어에 대한 놀라움 정보량을 계산하였다. 그 다음, 놀라움 정보량은 실제 단어 인식에 대한 반응 시간과 양의 상관이 있음을 규명하였다.

4. Linear Mixed Regression Model

단어 정보량 즉 놀라움 측정값이 시각적 단어 인식의 반응 시간과 선형 상관관계가 있는지를 모델링하기 위해 선형 혼합 회귀 분석을 사용하였다. 이 회귀 모형은 하나 이상의 독립 변수에 대해 반응 변수 간의 선형 관계를 설명한다.

선형 혼합 회귀 모형은 고정 효과와 임의 효과의 두 부분으로 구성된다. 고정 효과 항은 일반적으로 기존 선형

회귀 부분 인 반면, 임의 효과는 모집단에서 임의로 추출된 개별 실험 단위와 연관된다.

파생단어에 대한 시각 반응 시간은 각 학습자마다 다르기 때문에 임의 효과로 처리했으며, 고정 효과는 Morfessor 비지도 학습을 통해서 얻은 파생단어에 대한 놀라움 정보량으로 설정하였다. 본 논문에서는 Python의 Statsmodels 패키지를 사용하여 선형 혼합 효과 모델을 설계하였다.

III. The Proposed Language Model

3장에서는 한글 코퍼스의 단어 단위를 이용하여 d-형태소를 분석하는 언어 모델을 설계한다. 그리고 이 언어 모델의 단어 형태소 분석의 결과와 언어 심리 실험에서 얻은 한국어 언어사용자의 단어 이해 판단 시간의 상관관계에 대한 통계 선형 혼합 회귀 모형을 설계한다.

자연어 전처리 과정을 통하여 한국어 텍스트를 띄어쓰기를 기준으로 단어들로 분리하고, 형태소 분석에 필요한 학습 데이터, 전처리, 비지도 학습 방법, 그리고 학습한 모델의 테스트 방법을 기술한다.

1. Language Model Design

이 절에서는 언어 모델에 의해 파생단어를 처리 분석하는 과정을 설계한다.

그림 3에 제시된 바와 같이, 전처리 과정을 진행한 후 형태소 레이블이 없는 학습 데이터(Sejong Corpus)를 Morfessor Baseline모델에 입력하여 통계 기반 언어 모델을 훈련시킨다(①). 테스트 데이터(Korean Lexicon Project(KLP) 코퍼스)를 언어 모델에 입력하여 d-형태소 파생단어를 예측한다(②). 단어 빈도수(Word frequency), 단어 길이(Word length), 단어 반응시간(Reaction times), 단어 아이디(Word order number), 단어의 놀라움 자질을 사용하여 선형 혼합 회귀 모형의 변수로 사용한다(③).

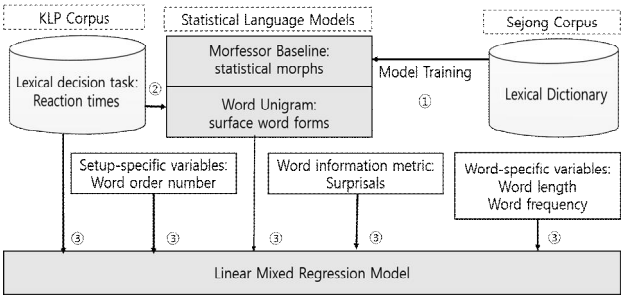

Fig. 3. The Workflow for Proposed Model

2. Training data

그림 3의 언어 모델을 학습 훈련하기 위해 국립국어원에서 제공하고 있는 세종 코퍼스를 사용하였다[6]. 본 실험에서는 3만 어절 이상을 갖고 있는 현대 문어 원시 파일들을 사용하였다. 표 1과 같이 각 파일은 “<!DOCTYPE>”시작하는 헤더가 구성되며, 문단은 로 구분된다.

Table 1. A Sample for Training Data Set

File
<!DOCTYPE tei.2 SYSTEM "c:\sgml\dtd\tei2.dtd" [<!ENTITY % TEI.corpus "INCLUDE"> <!ENTITY % TEI.extensions.ent SYSTEM "sejong1.ent"> <!ENTITY % TEI.extensions.dtd SYSTEM "sejong1.dtd">]>
…
지금 도처에서 '문화'의 바람이 불고 있다. 20세기 후반부터 그리고 21세기에 들어서면서부터…   그렇다면 왜 하필 지금 '문화'인가? 문화는 인류 역사와 더불어 늘 있어왔는데, …   그렇다면 왜 하필 지금 '문화'인가? 문화는 인류 역사와 더불어 늘 있어왔는데, …

Morfessor 알고리즘이 학습한 훈련 데이터는 표 2와 같이 세종코퍼스에서 현대 문어 459파일을 내려 받아 전처리 과정에서 오류가 발생하는 문장을 제외하고 총 문장의 개수 1,780,712에서 약 1백 2십만 개 어절을 학습 데이터로 선정하여 학습을 시켰다. 본 실험에서는 코퍼스를 랜덤하게 섞어 90%를 학습용 데이터로, 남은 10%를 검증용 테스트 데이터로 구성하였다.

Table 2. Training Data Extracted from Sejong Corpus

Item	Total number
Files	459
Words with 2~3 syllables	420043
Words	1260130
Sentences	1780712

구축한 언어 모델이 실제 어느 정도로 d-형태소 분석을 하는지를 테스트하기 위해 전문가에 의해 정교하게 각 단어별 형태소로 분석한 결과물을 테스트 데이터로 설정하였다. 테스트 데이터에서 무작위로 1000개의 샘플을 추출하여 Morfessor 알고리즘을 테스트할 때 준지도 검증 데이터로 사용하였다. 표 3은 전문가에 의해 테스트 데이터 각각에 대해 토큰으로 분리한 표이다.

Table 3. A Sample for Annotated Test Data

Lexicion	Tokenization
가겟집	가겟 + 집
가격표	가격 + 표
가계부	가계 + 비
가깝다	가깝 + 다

또한, 파생단어 판단 시간을 수집하기 위해 한국어 심성어휘집(Korean Lexicon Project: KLP) 어휘 판단 데이터베이스를 사용하였다. 한국어 단어 30,930개에 대한 어휘 판단 시간을 내려 받아 사용하였다. 이 프로젝트의 실험 참가자들은 국내 소재 4개 종합대학에서 최종적으로 참가한 52명이었고 이들의 평균 연령은 21.9세(범위: 18세-25세)이었고, 성별은 여자 29명 남자 23명이였다. 각 참가자는 화면 중앙에 제시되는 자극이 단어인지 혹은 비단어인지 판단하는 과제에서 각 단어의 반응 시간이 측정되었다[12-15].

표 4는 KLP 데이터베이스에 등록된 참가자들에게서 수집한 샘플 단어들의 어휘 판단 시간(Stim_RT)과 어휘 빈도수(Freq)를 포함하며, 또 동일한 어휘에 대해 전문가가 태깅한 형태소 분석에 대한 결과(Stim_val, match/mismatch)를 포함한다. 첫 번째 축은 파생단어, 두 번째 축은 전문가에 의해 측정한 d-형태소 분석 정보, 세 번째 축은 한국어 심성어휘집 어휘 판단 데이터베이스에서 단어 빈도수, 네 번째 축은 Morfessor 알고리즘에 의한 d-형태소 분리한 결과와 전문가가 형태소 분리한 결과가 일치 여부를 표시하며, 다섯 번째 축은 어휘 판단 시간이다.

Table 4. A Sample for Words with Response Time

Stimuli	Stim_val	Freq	match/ mismatch	Stim_RT
가건물	2.3	33	match	736.811
가르침	1	460	mismatch	542.174

3. Training Data Preprocessing

자연어 학습 데이터를 기계학습 모델에 학습시키기 이전에 문장에 포함되어 있는 불필요한 특수문자 제거 등과 같은 전처리가 필요하다. 각 문장들을 자연어 처리에 사용되는 Python 패키지 중 하나인 NLTK을 이용하여 단어별로 분리하였다.

학습 모델의 전처리 과정에서 다음과 같은 사항을 고려하였다. 첫 번째, Morfessor 알고리즘에 의해 각 한국어 단어에 대해 d-형태소로 구성된 결과를 추출하기 때문에 별도로 한국어 기반 품사 태깅을 진행하지 않았다. 두 번째, 하나 이상의 문장으로 구성된 문단들을 고려하였다. 세 번째, “.”, “ : ”, “ / ”, “@” 등과 같은 특수 기호 및 한자 및 영어를 제거하였다.

세종코퍼스에서 내려 받은 텍스트 파일을 대상으로 전처리 작업을 수행한 후 텍스트 파일에서 출현한 단어의 빈도수를 계산하여 구축한 그림 1의 어휘 사전에 대한 예시는 표 5와 같다. 표 5의 첫 번째 축은 단어의 빈도수, 두 번째 축은 추출한 단어이다.

Table 5. A sample for Lexicon Dictionary

Lexicion	Word Frequency
있다	124136
있는	122093
나는	56603
같은	44559

4. Unsupervised Machine Learning Training

모델 학습 과정에서는 전처리한 학습 데이터 기반으로 비지도 학습 모델을 앞서 제시한 그림 1과 같이 학습시킨다. 어휘사전을 Morfessor 학습 모델에 입력한다. Morfessor 알고리즘의 MDL 목적 함수의 비용이 최소화되도록 optimizer를 통해 Morfessor의 가중치를 수정하며 학습을 진행한다. 표 6은 학습한 모델에서 d-형태소 분리 규칙 유형 일부를 추출한 결과 예시이다.

Table 6. A Sample for Construction Types

Construction Types
('ㄸ가', 1), ('ㄸ문에', 1), ('ㄻ은', 1), ('ㅅ같이', 1), ('ㅅ거긔', 1), ('ㅅ건설', 1), ('ㅅ곱이', 1), ('ㅅ과', 1), ('ㅅ그룹', 1), ('ㅅ꼴로', 1), ('ㅅ대', 1), ('ㅅ대학', 1), ('ㅉ차', 1), ('ㅌ사', 1), ('ㅌ사는', 1), ('ㅌ사의', 1), ('ㅌ으로', 1), ('ㅌ은', 1), ('ㅌ의', 1), ('ㅌ이', 1), ('ㅌ티읕', 1), ('ㅐ로', 1), ('ㅐ애', 1), ('ㅐ와', 1), ('ㅐ외', 1), ('ㅑ ㅓ', 1), ('ㅑ길', 1), ('ㅑ야', 1), ('ㅒ애', 1), ('ㅒ와', 1), ('ㅓ나', 1), ('ㅓ로', 1), ('ㅓ어', 1), ('ㅖ로', 1), ('ㅖ로도', 1), ('ㅖ에', 1), ('ㅕ로', 1), ('ㅕ여', 1), ('ㅖ는', 1), ('ㅖ로', 1), ('ㅖ를', 1), ('ㅖ예', 1), ('ㅗ가', 1), ('ㅗ고', 1), ('ㅗ오', 1), ('ㅗ와', 1), ('ㅗ일', 1), ('ㅡㅣ로', 1), ('ㅡ가', 1), ('ㅡ년여', 1), ('ㅡ로', 1), ('ㅡ으', 1), ('ㅡ의', 1), ('ㅡ자', 1), ('ㅡ자집', 1), ('ㅢ는', 1)
...

본 실험에서는 d-형태소 분리 규칙을 사용하여 첫 번째 규칙의 개수(# of construction)를 기반으로 학습을 시킨 Morfessor모델 성능과 두 번째는 규칙의 유형 개수(# of construction types)를 기반으로 학습을 시킨 Morfessor 모델의 성능을 비교하였다. 첫 번째 Morfessor 모델이 훨씬 정확도(99%이상)가 높았다.

Table 7. Comparison of Morfessor Model Training

Performance	# of Constructions	# of Construction Types
F-score	0.998	0.692
Precision	0.996	0.529
Recall	1.0	1.0

5. Correlation between Surprisal and RT

그림 3에서처럼, Morfessor 형태소 분석 모델에 의해 테스트한 단어들 중 3음절로 구성된 파생단어를 입력하였을 때, 파생단어에 대한 어휘 판단 시간의 예측은 다음 세 단계로 분석되어진다.

첫 번째는 최적 Morfessor 언어모델을 사용하여 입력 단어에 대해 예측한 결과 d-형태소 시퀀스를 준비한다. 두 번째는 동일한 단어를 전문가에 의해 태깅한 형태소 시퀀스를 준비한다. 세 번째는 Morfessor에 의한 d-형태소 분석 결과와 태깅한 d-형태소 분석 결과가 일치하는 단어들만 대상으로 실험자가 시각적으로 단어를 읽었을 때 반응시간과 단어의 놀라움 측정치 간의 상관관계를 분석한다.

본 논문에서는 Python 언어의 Statsmodels 패키지에서 제공하고 있는 regression.mixed_linear_model 클래스를 사용하여 선형 혼합 회귀 모델을 다음과 같이 설계하였다.

$$model_{2-1} = MixedLM.from\_formula(\text{RT~surprisal*Freq}, groups = data['\text{words}'], data = data_{2-1})$$

$$model_{1-2} = MixedLM.from\_formula(\text{RT~surprisal*Freq}, groups = data['\text{words}'], data = data_{1-2})$$

Fig. 4. Linear Mixed Regression Model

여기서, groups=data[“material]라는 기호는 실제 실험 자료에 대한 랜덤 효과 변수를 의미한다. RT 기호는 단어에 대한 반응 시간 값의 총합을 구해 평균을 구하고 표준화한 값을 의미한다. surprisal 기호는 놀라움 변수를 의미한다. 모델 $model_{2-1}$은 하나의 종속 변수 RT와 고정효과 2 개의 변수 surprisal과 Freq를 가진 선형 회귀 함수($RT \sim surprisal*Freq$)로서 파생단어가 2개의 음절과 1개의 음절로 분리되는 파생단어($data = data_{2-1}$)를 대상으로 한 선형 혼합효과 모델을 나타낸다. 모델 $model_{1-2}$은 1개의 음절과 2개의 음절로 분리되는 파생단어($data = data_{1-2}$)를 대상으로 설계된 선형 혼합 효과 모델을 나타낸다.

IV. Experiment

4장에서는 데이터 기반 형태소 분석 언어 모델을 통해 파생단어 처리한 결과를 분석한다. 시스템 구현을 위해 구축한 실험 환경과 결과를 기술한다.

1. Hardware and Software

실험에서 사용한 하드웨어는 표 8과 같다. 학습 시간을 단축하기 위해 GPU의 사양 NVIDA GTX 1080을 사용하였다. CPU는 i5-2500K이며, Memory는 32G로 구성하였다. HDD는 256G SSD를 사용하였다.

Table 8. Hardware Configuration

Name	Version
GPU	NVIDA GTX 1080
CPU	i5-2500K
Memory	32G
HDD	256G SSD

실험에서 사용한 소프트웨어는 표 9과 같다. Python 기반 프로젝트 진행을 위해 Python 개발 툴킷인 PyCharm을 이용하였다. 자연어 처리 툴킷인 NLTK 라이브러리를 이용하여 전처리를 진행하였다. Morfessor 모듈은 Morfessor 2.0 패키지를 이용하였다.

Table 9. Software Configuration

Name	Version
NLTK	3.2.5
Morfessor	2.0
statsmodels	1.1-19

2. Results

파생단어에 대한 놀라움 변수과 빈도수 변수를 사용하여 단어 판단 시간을 예측하는 모델을 변수 간 교호작용을 고려하지 않은 모델과 고려한 모델로 실험하였다. 교호작용은 한 변수에 의해 다른 변수의 효과가 변하는 것을 의미한다.

그림 4의 선형 혼합 회귀 모델의 놀라움과 빈도수 요인의 교호작용 효과가 단어 판단 반응시간을 통계적으로 유의미하게 변화시킬 수 있었다. 또한 Morfessor모델의 파생단어에 대한 형태소 분석 결과와 전문가의 파생단어에 대한 형태소 분석 결과가 일치한 파생 단어와 불일치하는 단어를 분리하였다. 분석 결과가 일치하는 파생단어 중 음절 3개로 구성된 단어에 대해 2-1음절로 d-형태소 분석한 단어는 4023개, 1-2음절로 d-형태소 분석한 단어는 1137개이었다.

그림 5에서 두 개의 형태로 d-형태소 분리되었을 때 파생단어의 빈도수와 평균 반응 시간 사이에 음의 상관관계가 있음을 확인할 수 있다. 즉, 단어의 빈도수가 높을수록 단어 판단 반응시간이 짧아진다는 것은 경험상 익숙한 단

어를 인식하는데 걸리는 시간은 오래 걸리지 않는다고 해석할 수 있다. 1-2음절로 분리된 파생단어의 개수는 상대적으로 2-1음절로 분리된 파생단어의 개수보다 훨씬 적어서 빈도수와 단어 판단 반응 시간 사이의 회귀선이 가파르게 감소함을 알 수 있다.

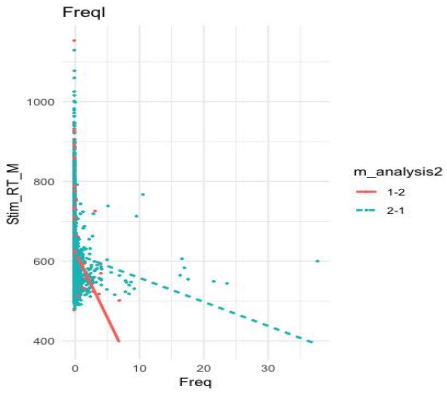

Fig. 5. Plots with Freq & Stim_RT_M with Segmentations

표 10은 Morfessor의 형태소 분석 결과와 전문가가 직접 태깅한 결과가 일치한 단어들 중 2개의 음절과 1개의 음절로 형태소 분리한 데이터를 일부 추출한 표이다.

Table 10. Words with 2-1 Segmentation

Word	Surprisal	RT	Freq
가격+표	49.021	2.3	17
가계+부	47.650	2.3	64
가공+품	49.435	2.7	29
가늘+다	46.973	2.3	653
가담+자	47.259	2.3	28
가락+지	46.858	2.3	17
가래+침	49.403	2.3	30
가로+등	48.875	2.3	284
가마+솥	52.015	2.3	61
가면+극	49.542	2.3	17
가부+좌	49.775	2.3	15

그림 6과 표 111은 표 10의 파생단어를 대상으로 선형 혼합 회귀 분석 결과이다. 즉, 파생단어의 놀라움과 판단 반응시간은 양의 상관관계가 있으며 이는 통계적으로 유의미한 것을 알 수 있다(p value=0.000 < 0.05). 그러나 단어의 빈도수는 판단 반응 시간과 상관이 없으며, 놀라움과 빈도수의 교호작용 역시 단어 판단반응 시간에 효과가 없음을 알 수 있다. 따라서 2-1 음절로 구성된 파생단어에 대한 놀라움 정도가 높을수록 단어에 대한 판단/이해 반응 시간은 길어진다고 해석할 수 있다.

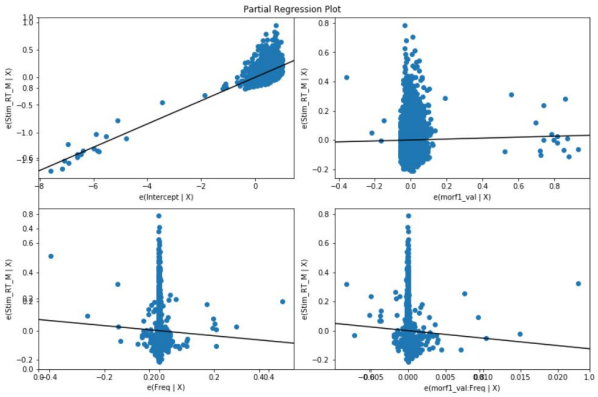

Fig. 6. Plots with 2-1 Segmentation

Table 11. LM Regression: 2-1 Segmentation

	estimate	std.err	z value	pr(>|z|)
Interc	0.211	0.002	104.12	0.000
Surpr	0.032	0.006	5.411	0.000
Freq	-0.174	0.118	-1.468	0.142
Surpr:Freq	-5.100	2.900	-1.759	0.079

표 12은 Morfessor의 형태소 분석결과와 전문가가 직접 태깅한 결과가 일치한 파생단어들 중 1개의 음절과 2개의 음절로 형태소 분리한 단어를 일부 추출한 표이다.

Table 12. Words with 1-2 Segmentation

Word	Surprisal	RT	Freq
가+건물	46.605	2.7	33
가+계약	46.605	2.7	19
가+등기	46.605	2.7	16
가+석방	46.605	2.7	49
가+압류	46.605	2.7	29
가+처분	46.605	2.3	52
간+세포	48.358	2.7	46
감+나무	48.509	2.7	190
값+어치	50.276	2.7	46
강+바닥	48.501	2.7	30
강+바람	48.501	2.7	19

그림 7과 표 13은 표 12의 1개 음절과 2개의 음절로 형태소 분리한 파생단어에 대해 선형 혼합 회귀 분석 결과이다. 여기서는, 파생단어의 놀라움과 판단 반응시간은 양의 상관관계가 있지만 통계적으로 미미하게 유의미한 것을 알 수 있다(p-value=0.054). 한편, 단어의 빈도수와 반응 시간이 음의 상관이 있으며(p-value < 0.05), 놀라움과 빈도수의 교호작용 효과는 단어 반응 시간에 효과가 있음을 알 수 있다(p-value=0.021 < 0.05). 따라서 1-2 음절로 구성된 파생단어에 대한 놀라움 정도가 높을수록 단어에 대한 판단/이해 반응 시간은 길어진다고 해석할 수 있다. 또한 2-1 음절로

구성된 파생단어에 대한 놀라움 정도(0.032)보다 1-2 음절로 구성된 파생단어에 대한 놀라움 정도(0.037)가 평균적으로 약간 높은 이유는 실제로 음절 3개로 구성된 파생단어에서 2개의 음절을 보고 마지막 음절을 예측하는 것보다 첫 번째 음절을 보고 다음 두 개의 음절을 예측하는 것이 상대적으로 어렵다는 점을 보여준다.

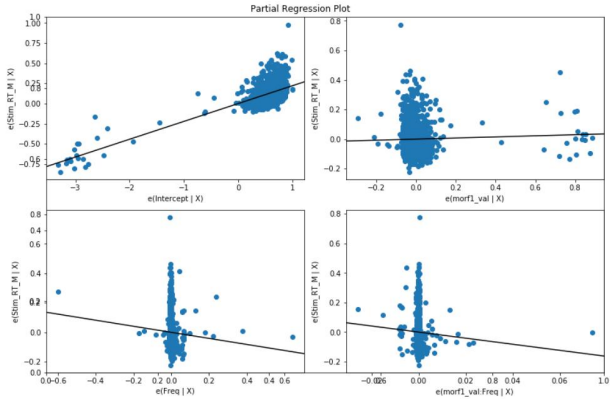

Fig. 7. Plots with 1-2 Segmentation

Table 13. LM Regression: 1-2 Segmentation

	estimate	std.err	z value	pr(>|z|)
Interc	0.221	0.003	71.323	0.000
Surpr	0.037	0.019	1.923	0.054
Freq	-0.205	0.065	-3.156	0.002
Surpr:Freq	-2.056	0.889	-2.311	0.021

V. Conclusions

한국어 단어 코퍼스를 대상으로 학습시킨 비지도 기계학습 모델을 이용하여 파생단어 인식을 예측하는 연구는 현재까지 찾아보기가 쉽지 않다. 본 논문에서는 비지도 기계학습 기술과 코퍼스의 단어 단위를 이용하여 한국어 단어를 형태소 분석하는 언어 모델을 구축하였다. 제안한 시스템은 언어 모델과 선형회귀 모델로 구성되어진다.

첫 번째 세종 코퍼스의 일부를 대상으로 Morfessor 기반 언어 모델을 이용하여 3음절로 구성된 파생단어를 2가지 형태 구조로 분리한 d-형태소 분리 분석을 수행하였다. 1,780,712개의 문장을 띄어쓰기 기준으로 전처리 후 420,043개 단어로 토큰화해서 Morfessor 모델로 학습을 시켰다. 이 모델은 순수하게 토큰만으로 비지도 학습을 했으며, 단어에 대한 2음절과 1음절 또는 1음절과 2음절로 구성된 형태소 분석 분류를 예측하였다.

두 번째 선형 혼합 회귀모델을 이용하여 형태소 분석 기반 언어 모델의 파생단어의 놀라움 정도와, 어휘 판단 과제 실험을 통해 동일 단어를 읽는데 걸리는 반응 시간과의 상관관계를 분석하였다.

실험 결과는 다음과 같이 해석할 수 있다. 제안된 비지도 기계학습의 언어 모델은 파생단어를 d-형태소로 분석해서 파생단어의 음절의 형태로 처리를 하였다. 파생단어를 처리하는데 필요한 사람의 인지 노력의 양 즉, 판독 시간 효과가 실제로 형태소 분류하는 기계학습 모델에 의한 단어 처리/이해로부터 초래될 수 있는 놀라움과 상관함을 보여 주었다. 본 연구는 놀라움의 가설 즉, 놀라움 효과는 단어 읽기 또는 처리 인지 노력과 관련이 있다는 가설을 뒷받침함을 확인하였다.

본 연구는 3음절 단어의 형태 분석에 초점을 맞추었다. 이 논문의 결과가 2음절 혹은 4음절 이상의 한국어 단어에도 제안한 언어 모델에 적용했을 때 유의미한지는 향후 과제로 남긴다. 또한 본 연구를 통해 Morfessor의 형태 분석의 효과성이 상당히 높다는 것을 확인하게 되었지만, 언어 전문가의 형태 분석 수준에는 아직 도달하지 못하고 있다. 이를 개선하기 위해 단어의 품사 태깅, 접미사와 접두사 태깅을 통해 준지도 학습 기반 모델을 구축하는 연구는 앞으로 수행할 과제로 남긴다.

ACKNOWLEDGEMENT

This work was supported by 2019 Shinhan Univ. Research Grant.

REFERENCES

[1] A. J and M. O, “What Your Username Says About You,” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2302-2307, Sept. 2015.

[2] M. Creutz, and K. Lagus, “Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor1.0,” Helsinki University of Technology, March 2006.

[3] S. Virpioja, M. Lehtonen, A. Hulten, H. Kivikari, R. Salmelin, and K. Lagus, “Using Statististical Models of Morphology in the Search for Optimal Units of Representation in the Human Mental Lexicon,” Cognitive Science, Vol. 42, pp. 939-973, March 2018.

[4] M. Lehtonen, M. Varjokallio, H. Kivikari, A. Hulten, S. Virpioja, T. Hakala, M. Kurimo, K. Lagus, and R. Salmelin, “Statistical models of morphology predict eye-tracking measures during visual word recognition,” Memory&Cognition, Vol. 47, Issue 7, pp. 1245-1269, May 2019.

[5] G. Booij. “The Grammar of Words: An Introduction to Linguistic Morphology. Oxford Textbooks in Linguistics,” OUP Oxford, Sept. 2012.
[6] Sejong-Corpus, http://ithub.korean.go.kr/user/main.do
[7] Kkma, http://kkma.snu.ac.kr/documents/?doc=postag
[8] UTagger, http://nlplab.ulsan.ac.kr/doku.php?id=utagger
[9] Khaiii, https://tech.kakao.com/2018/12/13/khaiii/
[10] S. Virpioja, P. Smit, S-A. Gronroos, and M. Kronroos, “Morfessor 2.0: Python Implementation and Extensions for Morfessor Baseline,” Technical Report, Aalto University publication series SCIENCE + TECHNOLOGY, 25, pp. 38, Dec. 2013.
[11] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.” IEEE Transactions on Information Theory, 13(2):260−269, April, 1967
[12] Korean Lexicon Project, http://klexicon.org
[13] K. Yi, M-M. Koo, K. Nam, K. Park, T. Park, S. Bae, C-H. Lee, H-W. Lee and J-R. Cho, “The Korean Lexicon Project-A Lexical Decision Study on 30,930 Korean Words and Nonwords,” The Korean Journal of Cognitive and Biological Psychology, pp. 395-410, Oct. 2017.
[14] E. Kim, “A Deep Learning-based Article- and Paragraph-level Classification,” The Journal of the Korea Society of Computer and Information, pp. 31-41, Nov. 2018.
[15] J. Park, A. Seok, Y. Yoon, and B. Rhee, “An Analysis of Instagram Hashtags Related to the Exhibitions in Korean,” The Journal of the Korea Society of Computer and Information, pp. 49-56, March 2019.

Authors

Euhee Kim received the M.S. degrees in Computer Engineering from Dongguk University, Korea, in 2002 and Ph.D. degree in Mathematics from The University of Connecticut, U.S.A in 1995.

Dr. Kim is currently an associate Professor in the Department of Computer Science & Engineering at Shinhan University. She is interested in AI and Big Data computing.