JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 12, pp. 145-153, December 2020
https://doi.org/10.9708/jksci.2020.25.12.145

Interactive Morphological Analysis to Improve Accuracy of Keyword Extraction Based on Cohesion Scoring

Yang Woo Yu*, Hyeon Gyu Kim**

*Associate Professor, Dept. of Digital Contents Design, Ulsan College, Ulsan, Korea
**Associate Professor, Div. of Computer Science and Engineering, Sahmyook University, Seoul, Korea

[Abstract]

Recently, keyword extraction from social big data has been widely used for the purpose of extracting opinions or complaints from the user's perspective. Regarding this, our previous work suggested a method to improve accuracy of keyword extraction based on the notion of cohesion scoring, but its accuracy can be degraded when the number of input reviews is relatively small. This paper presents a method to resolve this issue by applying simplified morphological analysis as a postprocessing step to extracted keywords generated from the algorithm discussed in the previous work. The proposed method enables to add analysis rules necessary to process input data incrementally whenever new data arrives, which leads to reduction of a dictionary size and improvement of analysis efficiency. In addition, an interactive rule adder is provided to minimize efforts to add new rules. To verify performance of the proposed method, experiments were conducted based on real social reviews collected from online, where the results showed that error ratio was reduced from 10% to 1% by applying our method and it took 450 milliseconds to process 5,000 reviews, which means that keyword extraction can be performed in a timely manner in the proposed method.

▸ **Key words**: Big data, Social reviews, Keyword extraction, Cohesion score, Interactive morphological analysis

[요 약]

최근 소셜 빅데이터를 대상으로 한 키워드 분석은 고객 관점의 의견이나 불만 사항을 추출하기 위한 목적으로 광범위하게 활용되고 있다. 이와 관련하여, 이전 연구에서는 키워드 분석의 정확도를 높이기 위해 응집도 점수를 활용한 방법을 제안하였으나, 리뷰의 수가 적을 경우 오류율이 증가하는 문제가 있었다. 본 논문에서는 응집도 점수 기반 알고리즘으로부터 추출된 키워드에 대해 간소화된 형태소 분석 단계를 후처리 형태로 적용함으로써 키워드 추출의 정확도를 개선하고자 하였다. 제안 방법은 입력 데이터가 주어질 때마다 필요한 형태소 분석 규칙을 점증적으로 추가할 수 있도록 지원함으로써, 사전의 크기를 최소화하고 분석의 효율을 높이고자 하였다. 또한 대화형 규칙 입력 시스템을 제공하여 분석 규칙 추가에 드는 노력을 최소화하고자 하였다. 제안 방법을 검증하기 위해 온라인에서 수집된 실제 리뷰를 대상으로 실험을 수행하였으며, 제안 방법을 적용할 경우 오류율이 기존 10%에서 1%로 개선되는 동시에, 5,000개의 리뷰 처리에 450ms가 소요되어 실시간 처리가 가능한 수준임을 확인하였다.

▸ **주제어**: 빅데이터, 소셜 리뷰, 키워드 추출, 응집도 점수, 대화형 형태소 분석

• First Author: Yang Woo Yu, Corresponding Author: Hyeon Gyu Kim
*Yang Woo Yu (soft@uc.ac.kr), Dept. of Digital Contents Design, Ulsan College
**Hyeon Gyu Kim (hgkim@syu.ac.kr), Div. of Computer Science and Engineering, Sahmyook University
• Received: 2020. 10. 26, Revised: 2020. 11. 24, Accepted: 2020. 11. 24.

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

I. Introduction

최근 들어, 고객 관점의 의견이나 불만 사항을 추출하기 위한 목적으로, SNS 피드 및 블로그 리뷰 등을 포함한 소셜 빅데이터가 대중적으로 활용되고 있다[1, 2]. 소셜 빅데이터는 네이버 및 구글 등의 온라인 포털 업체에서 제공하는 오픈 API[3, 4]를 통해 무료로 획득 가능하며, 신용카드, 휴대폰 이용 내역 등 수치로 이루어진 다른 형태의 빅데이터에 비해 고객들의 의견이나 불만 사항을 텍스트 형식으로 바로 파악할 수 있다는 점에서 선호되고 있다.

소셜 빅데이터 분석에서는 형태소 분석을 통한 키워드 추출 과정이 필수적으로 요구된다. 예를 들어, “화덕피자로 엄청 유명해진 그라파피자리아, 인싸들의 핫플!”이라는 리뷰를 가정해보자. 먼저 형태소 분석을 통해 각 단어의 원형을 추출해 내는 작업이 선행되며, 다음으로 의미가 크지 않은 부사나 동사 등의 제거 작업이 이루어진다. 예를 들어, “유명해진”은 “유명하다”의 원형 형태나 “유명한” 등 대중으로부터 가장 많이 언급된 일반 형태로 변경되며, 의미가 크지 않은 “엄청” 등의 부사는 제거된다. 따라서 주어진 리뷰로부터 {“화덕피자”, “유명한”, “그라파피자리아”, “인싸”, “핫플”} 등을 포함한 단어 집합이 키워드 추출 결과로 반환된다.

위 예제에서도 볼 수 있듯이, 소셜 리뷰는 고유명사(“그라파피자리아”)나 트렌드를 반영한 신조어(“인싸”, “핫플”)를 포함하는 경우가 많다. 문제는 기존 형태소 분석기들이 이를 잘 처리하지 못한다는 점이다. 그 이유는 기존 분석기들의 경우 형태소 분석을 위해 단어 사전을 이용하나, 해당 사전에는 고유명사나 신조어가 포함되지 않아 관련 명사 추출이 어렵다는 점에 기인한다. 예를 들어 현재 대중적으로 이용되고 있는 형태소 분석기 중 하나인 꼬꼬마[5]의 경우, 위 리뷰에 대해 {“화덕”, “피자”, “그”, “라파”, “피자”, “리아”, ...} 등의 형태로, 사전에 포함된 단어 위주로 단어를 세분화하여 결과를 전달한다. 실제 리뷰에서 중요한 의미들이 주로 명사에 포함된다는 점을 고려할 때, 고유명사 및 신조어 추출의 정확도를 높이는 작업은 필수적이라 할 수 있다.

위 문제를 해결하기 위해, 이전 연구[6]에서 단어별 빈도수를 기반으로 계산된 응집도 점수를 활용한 통계적인 키워드 추출 방법을 제안하였다. 그리고 해당 방법이 신조어나 고유명사 처리에 있어 기존 방법에 비해 높은 정확도를 나타내는 동시에, 실시간 처리가 가능함을 입증하였다. 이에 반해, 응집도 점수 계산에 필요한 데이터가 충분하지 못할 경우, 형용사나 동사 형식의 활용형 단어에 대한 원형 처리가 이루어지지 않거나, 명사로부터 조사가 제대로 분리되지 않는 등, 다소 부정확한 결과가 제공되는 문제점이 있었다.

본 논문에서는 이전 연구의 문제점을 보완하기 위해 대화형 형태소 분석기를 활용한 방법을 제안한다. 제안 방법의 핵심은 응집도 점수 기반의 키워드 추출 결과로 얻어진 단어 집합에 대해 간소화된 형태소 분석 단계를 후처리의 형태로 추가/적용하는 것이다. 형태소 분석기는 대화형으로 구성하였으며, 입력 데이터가 추가될 때마다 각각의 패턴에 맞는 형태소 분석 규칙을 점증적으로 추가할 수 있도록 구현하였다. 이러한 구조를 통해, 키워드 추출에 활용되는 단어와 변환 규칙을 최소화하여 형태소 분석의 효율을 높임으로써, 실시간 처리를 지원하는데 초점을 두었다.

한편, 대화형 시스템은 필연적으로 수작업을 수반한다. 예를 들어, 제안하는 방법에서는 입력 데이터가 주어질 때마다 단어별로 오류 여부를 확인하고, 필요한 경우 단어에 대한 형태소 변환 규칙을 수작업으로 추가해야 한다. 따라서 입력 데이터가 커질수록 규칙 추가 비용이 증가하는 구조이다. 이를 해결하기 위해 제안 방법에서는 규칙 추가를 반자동화한 시스템을 제공하고 있으며, 해당 시스템의 성능을 검증하기 위한 실험 결과, 온라인에서 수집된 약 6백만 개의 소셜 리뷰에 대한 형태소 변환 규칙을 완성하는데 약 36시간 정도 소요되는 것으로 파악되었다.

II. Related Work

앞서 언급한 바와 같이, 기존의 형태소 분석 알고리즘들은 사전을 기반으로 분석을 수행한다. 따라서 다수의 연구에서 사전의 구조나 검색 방법을 최적화하여 효율을 높이기 위한 방법을 제안하였다. 대표적으로, [7]에서는 사전 탐색 회수를 줄이기 위한 형태소 분석 알고리즘을 소개하였으며, [8]에서는 사전의 구조를 개선하여 효율을 높이고자 하였다. 또한 [9]에서는 자주 이용되는 어절에 대한 분석 사전을 미리 구축해 두고, 이를 기반으로 코드 변환, 원형 복원 등의 절차적인 분석 단계를 생략하여 효율을 높이기 위한 방법을 소개하였다.

소셜 빅데이터 분석을 이용하는 대다수의 응용에서 역시 전통적인 형태소 분석 방법을 이용하여 키워드 추출을 구현하고 있다. [10]에서는 소셜 리뷰로부터 잠재적인 광고 키워드를 추출하였으며, [11, 12]에서는 영화 흥행 요인 분석을 위해 소셜 빅데이터를 이용하였다. 그리고 [13]에서는 온라인 쇼핑몰의 상품평 분류를 통한 감성 분석을 주제로

다루었다. 이들 응용에서는 KAIST의 한나눔 형태소 분석기[14]를 활용하여 소셜 리뷰로부터 키워드 집합을 얻는다고 가정하고 논의를 진행하였다는 측면에서 공통점이 있다.

최근 들어, 소셜 빅데이터 분석에 대한 관심이 높아지면서, 리뷰에 포함된 신조어나 고유 명사 등을 보다 정확하게 추출하기 위한 기법들이 제안되고 있다. 이들은 키워드 추출에 있어 사전이 아닌, 단어의 빈도수 등을 이용한 통계적인 분석 기법에 해당하며, 대표적으로 브랜칭 엔트로피(Branching Entropy)[15], 응집도 점수 산정 기법[16, 17, 18] 등이 포함된다. 이들은 공통적으로 단어 내부에서는 불확실성이 줄어들고, 단어의 경계에서는 불확실성이 증가한다는 점을 이용한다. 예를 들어, 주어진 단어가 “natur”까지 등장했다면, 그 다음에 나올 수 있는 글자는 “nature”이거나 “natural” 두 가지 경우에 국한되므로 불확실성이 낮아진다. 이에 반해 “nature”까지 등장할 경우, 다음 글자에 대한 불확실성이 다시 높아지므로, “nature”를 단어의 경계로 설정한다. 단, 경계를 판단하기 위한 산정식에 있어, 응집도 산정 기법에서는 엔트로피 값이 아닌 조건부 확률을 이용한다는 측면에서 차이가 있다. 이외에도 [19]에서는 비지도 기계학습과 코퍼스의 단어를 이용한 형태소 분석 방법에 대해 논의하였다.

III. Proposed Method

이 장에서는 이전 연구[6]에서 소개한 응집도 점수 기반 키워드 추출 알고리즘의 문제점을 보완하기 위한 대화형 형태소 분석기에 대해 소개한다. 제안하는 방법은 기존 분석기에 비해 간소화된 절차와 구조를 가지고 있으며, 이를 통해 키워드 추출의 정확도를 높이는 동시에 실시간 처리를 지원할 수 있도록 구성하였다.

1. Overview

제안하는 방법은 기존의 형태소 분석 방법에 비해 아래의 측면에서 차별화될 수 있다.

- 사전 구성 방법과 관련하여, 기존 방법은 형태소 분석을 위해 미리 완성된 사전을 이용하나, 제안 방법은 입력 데이터 집합이 주어질 때마다 사전에 규칙을 추가하는 점증적인 방식을 취함
- 형태소 분석 단위의 측면에서, 기존 방법은 자소 단위(예, ㄱ, ㄴ, ㅏ, ㅓ)로 분석을 진행하는 반면, 제안 방법은 음절(글자) 단위로 분석을 진행

제안 방법은 점증적인 사전 구성 방식을 채택하였으며, 사전에는 현재까지 주어진 입력 데이터의 처리에 필요한 형태소 분석 규칙과 단어만을 포함하므로 사전의 크기가 최소화된다. 또한 음절 단위로 분석을 수행하므로 자소 단위 분석을 수행하는 기존 방법에 비해 알고리즘의 구조가 크게 단순화된다(3.2절 참고). 따라서 제안 방법은 최소화된 형태소 변환 규칙과 단순한 분석 알고리즘을 바탕으로 분석 효율을 극대화한 구조로 볼 수 있다. 이를 통해 대량의 소셜 리뷰에 대한 실시간 형태소 분석 및 키워드 추출이 가능한 구조이다(4장 참고).

이에 반해, 제안 방법은 대화형 구조를 지니므로 필연적으로 수작업을 수반한다. 새로운 입력 단어 집합이 주어질 때마다 단어별로 형태소 변환 오류 여부를 확인하고, 필요한 경우 오류를 수정하기 위한 변환 규칙을 수작업으로 추가해야 한다. 따라서 형태소 변환 규칙 추가를 위한 수작업 노력을 줄이는 것이 무엇보다 중요하며, 제안 방법에서는 아래의 구조를 통해 문제를 해결하고자 하였다.

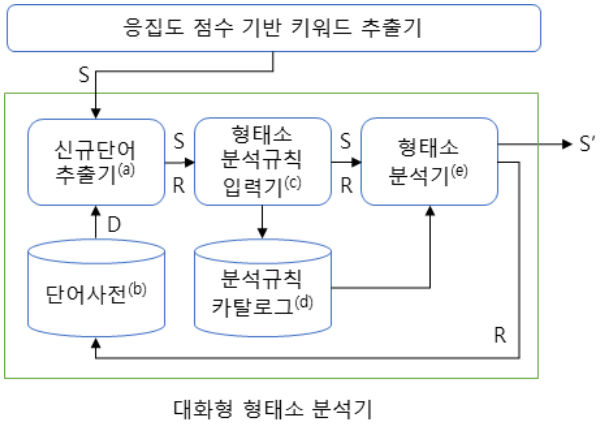

Fig. 1. Architecture of the proposed interactive morphological analyzer

앞서 언급한 바와 같이, 대화형 형태소 분석기는 이전 연구에서 구현된 응집도 점수 기반 키워드 추출기의 후처리 과정으로 이용되며, 추출된 키워드에 대해 간소화된 형태소 분석 과정을 적용시킴으로써 정확도를 개선하는 역할을 한다. 키워드 추출기로부터 생성된 단어 집합을 S라고 하자($S = \{w_i\}$, $1 \le i \le N$, w_i는 i번째 단어). [그림 1]에서 S는 가장 먼저 신규단어추출기(a)로 전달된다. 해당 모듈은 사전(b)에 포함된 단어 집합 D로부터 S를 제외한 차집합 R을 구한다($R = D - S = \{w_i\}$, $1 \le i \le K$, $K \le N$). 이를 통해 수작업으로 확인해야 하는 단어의 수를 대폭 줄인다.

신규 단어 집합 R은 형태소규칙입력기(c)로 전달된다. 해당 모듈은 먼저 R에 포함된 단어를 행단위로 나열한 엑셀

파일을 생성한다. 엑셀 파일의 각 행에는 단어에 오류가 있을 경우, 이를 보정하기 위한 형태소 변환 규칙과 변환에 필요한 추가 데이터를 정의하기 위한 컬럼을 제공한다([표 2] 참고). 관리자에 의해 엑셀 파일에 오류 수정을 위한 변환 규칙 정의가 완료되면, 형태소규칙입력기가 다시 해당 엑셀을 받아들여 단어별로 추가된 변환 규칙을 추출한 후, 이를 분석규칙카탈로그[(d)]에 추가한다.

분석규칙카탈로그의 내용이 업데이트되면, 이를 기반으로 형태소 분석기[(e)]가 업데이트된 규칙을 기반으로 입력 단어 집합 S에 대한 형태소 분석을 시작한다. 그리고 분석을 통해 보정된 키워드 집합 S'를 결과로 출력한다. 또한 R에 포함된 신규 단어를 단어사전[(b)]에 추가시킴으로써, 다음 입력 데이터가 주어졌을 때 확인해야 할 단어의 수를 점진적으로 줄여나갈 수 있도록 구성하였다.

2. Simplified Morphological Analysis

제안 방법의 형태소 분석은 추출된 키워드의 정확도를 높이기 위한 목적으로 간소화된 형태로 수행된다. 보다 자세히, 본 논문에서의 형태소 분석 범위는 다양한 활용 형태의 단어로부터 원형을 찾아내기 위한 것으로 국한되며, 단어별로 품사를 알아내기 위한 알고리즘은 포함하지 않는다. 이러한 가정을 바탕으로, [그림 1]의 형태소 분석기[(e)]는 아래의 변환 과정을 수행한다.

(1) 의미없는 단어 제외
(2) 명사일 경우 조사 제거
(3) 형용사/동사일 경우 활용 처리
(4) 유의어 처리

(1)은 입력 집합 R로부터 키워드로써 의미가 없는 중립적인 단어를 분석에서 제외시키는 과정이다. 예를 들어, "거의", "너무", "더욱" 등의 부사들은 형용사나 동사를 꾸미기 위한 단어에 해당하며, 키워드로는 큰 의미가 없다. 또한, "있다", "하다" 등의 일부 형용사나 동사의 경우 역시 의미 없이 형식적으로 이용될 수 있다. 이들 단어는 키워드로써 의미가 없는 경우가 많으므로, 형태소 분석 및 키워드 추천 과정에서 제외시킨다.

(2)는 명사를 얻기 위해 조사를 제거하는 과정이다. 조사로 구성된 단어 집합을 P라고 하자($P = \{p_i\}$, $1 \le i \le N_p$, p_i는 i번째 조사). 이 때, 입력으로 주어진 단어 x가 p_i로 끝난다면, $(x - p_i)$를 명사로 간주하고 리턴한다. 예를 들어, P = {이, 가, 은, 는}이라고 하자. 이 경우 입력 단어로 "여행이"가 주어질 경우, 해당 단어는 p_0 = "이"로 끝나므로 "여행"을 명사로 간주하고 반환한다.

단 일부 명사의 경우, 해당 단어 자체에 p_i와 동일한 표현이 포함될 수 있다. 예를 들어, 입력 단어가 "물놀이"나 "해돋이" 등으로 주어진다면, 위 방법에서는 "물놀"과 "해돋" 등이 결과로 반환된다. 이러한 문제는 간단한 예외 검사를 통해 해결가능하다. 예를 들어, 입력 단어가 "이"로 끝날 경우, 해당 단어가 "놀이"나 "돋이" 등으로 끝나지 않는지 체크한 후, 조건을 만족하는 경우에 한해 추출된 명사를 반환하는 형식을 취할 수 있다.

(3)은 형용사나 동사에 대한 원형을 구하기 위한 과정이다. 형용사 및 동사의 활용 처리 케이스는 매우 다양하므로 단기간에 이를 완벽하게 구현하기는 어렵다. 대신, 제안 방법에서는 간단한 패턴 매칭을 이용하여 원형을 구하고자 하였다. 예를 들어 "귀여", "귀엽" 등의 활용형 형태로 시작되는 단어가 주어질 경우, 원형인 "귀엽다"나 가장 대중적으로 언급되는 형태인 "귀여운" 등으로 대체할 수 있다. (본 논문에서는 편의상 후자를 원형과 같은 것으로 취급한다.) 동일한 방법으로 "넓게", "넓어" 등의 활용형은 "넓은" 등의 표현으로 대체 가능하다. 이 방법은 자소 단위의 복잡한 분석 없이도 간단한 패턴 매칭으로 구현 가능하며, 입력 데이터에 대한 검토 및 규칙 추가를 꾸준히 진행할 경우 정확도가 점진적으로 향상될 수 있다.

(4)는 유의어 처리를 위한 과정이다. 예를 들어, "초콜렛"과 "초콜릿"의 경우 같은 의미를 지님에도 불구하고 별도의 처리가 없을 경우, 두 개의 키워드가 따로 추천될 수 있다. 따라서 다양한 형태의 유의어에 대해 가장 일반적으로 언급되는 (리뷰에서 빈도수가 가장 높은) 대표 단어를 선정한 후, 이를 중심으로 유의어를 합병하는 작업이 필요하다. 위 예의 경우, "초콜렛"이나 "초컬릿" 등에 대해 가장 일반적인 형태인 "초콜릿"으로 변환시킨다.

Table 1. Catalog structure for the proposed morpho logical analysis

group	name	data example	size
(1)	S_{neut}	[가끔, 가장, 굳이, 그냥, ... 나름, 너무, 더욱, ...]	1,295
(2)	S_{post}	[가, 는, 도, 들, 로, ... 에, 와, 은, 을, 이, ...]	66
	S_{poex}	[같은, 같이, 구이, 놀이, ... 데이, 돋이, 레이, 린이, ...]	80
(3)	S_{conj}	[[귀여, 귀여운], [귀엽, 귀여운], [넓게, 넓은], [넓어, 넓은], ...]	246
(4)	S_{syn}	[[초컬릿, 초콜릿], [초콜렛, 초콜릿], ...]	27

[표 1]은 위에서 설명한 변환 과정을 구현하기 위해 필요한 분석규칙카탈로그[(d)]의 구조를 도식화하고 있다. 먼저

S_{neut}은 과정 (1)을 구현하기 위한 데이터 집합을 정의하고 있으며, 해당 집합에는 1,295개의 중립 단어들이 포함되어 있다. 해당 숫자는 온라인에서 수집된 약 6백만 개의 소셜 리뷰를 대상으로 추출된 결과를 제시하였다.

S_{post}와 S_{poex}는 과정 (2)를 구현하기 위한 데이터 집합으로, 각각 조사로 구성된 단어 집합과 조사와 동일한 단어로 끝나는 경우 예외 처리를 위한 단어 집합을 정의하고 있으며, 그 수는 각각 66개와 80개로 조사되었다. S_{conj}는 과정 (3)을 구현하기 위한 데이터 집합으로, 각 원소는 활용형과 원형을 포함한 두 개의 단어로 구성되어 있으며, 현재까지 246개 규칙이 파악되었다. S_{syn}은 과정 (4)를 구현하기 위한 데이터 집합으로, 각 원소는 유의어와 (이를 대체하기 위한) 대표어로 구성되어 있으며, 현재까지 27개의 규칙이 추가되었다.

[표 1]의 카탈로그 구조를 기반으로 주어진 단어에 대한 원형을 추출하기 위한 Java 알고리즘은 아래와 같이 기술되 수 있다. getRoot() 함수는 크게 네 개의 for 문으로 구성되어 있으며, 각각은 위에서 설명한 (1)부터 (4)까지의 처리 과정을 구현하고 있다.

```
function getRoot(word) {
    for (var i=0; i<neutwords.length; i++) {
        if (word.startsWith(neutwords[i])) return null;
    }
    for (var i=0; i<postpos.length; i++) {
        if (word.endsWith(postpos[i]) && !excPost(word)) {
            return word.substring(0, word.length -
                                  postpos[i].length);
        }
    }
    for (var i=0; i<conjwords.length; i++) {
        if (word.startsWith(conjwords[i][0])) {
            return conjwords[i][1];
        }
    }
    for (var i=0; i<synonyms.length; i++) {
        if (word == synonyms[i][0]) return synonyms[i][1];
    }
    return word;
}
```

Fig. 2. Java function to get a root word in the proposed method

첫 번째 for는 과정 (1)을 구현한 것으로, neutwords 변수는 S_{neut}를 나타낸다. 주어진 단어 word가 S_{neut}의 단어로 시작하면, word를 중립적인 단어로 판단하고 null을 리턴한다. 두 번째 for문은 과정 (2)를 구현한 것으로, postpos 변수는 S_{post}를 나타낸다. 해당 for 문에서는 if 문을 통해 주어진 word가 S_{post}의 단어로 끝나는 동시에, word가 예외 집합 S_{poex}의 단어로 끝나지 않는 경우, substring 함수를 이용하여 word에서 조사 부분(i.e., postpos[i])을 분리한 후 반환한다. 위 구현에서 excPost (word) 함수는 word가 S_{poex}의 단어로 끝나는 경우 true를 반환하며, 세부 구현 코드는 지면 관계상 생략하였다.

세 번째 for 문은 과정 (3)을 구현하였으며, conjwords 변수는 S_{conj}를 나타낸다. 만약 주어진 word가 S_{conj} 요소 중 하나(conjwords[i][0])로 시작된다면, 해당 요소의 원형값(conjwords[i][1])을 반환한다. 마지막 for 문은 과정 (4)를 구현한 것으로, synonyms 변수는 S_{syn}을 나타낸다. 만약 word가 S_{syn}의 유의어 요소 중 하나(synonyms[i][0])와 일치하면, 해당 요소의 대표어(synonyms[i][1])를 반환한다.

3. Interactive Rule Adder

제안 방법에서 [표 1]의 카탈로그 데이터는 형태소규칙 입력기[c]를 통해 추가된다. 해당 모듈은 규칙 추가의 편의를 제공하기 위해, 신규 단어 집합 R에 포함된 단어를 행 단위로 나열한 엑셀 파일을 생성/제공한다. 엑셀 파일의 각 행에는 단어에 오류가 있을 경우 이를 보정하기 위한 변환 규칙 번호와, 변환에 필요한 추가 데이터를 함께 정의하기 위한 컬럼을 제공한다.

Table 2. Example of rule definition for the proposed morphological analysis

word	type	val1	val2
벌써	1		
월요일은	2	은	
해돋이	2	이	돋이
귀여워	3	귀여	귀여운
초콜렛	4	초콜릿	

[표 2]는 형태소 분석 규칙 정의를 위한 엑셀 파일의 구성과 입력 예를 보여준다. type 컬럼에는 단어의 변환에 필요한 규칙 번호를 지정하며, 이전 절에서 정의한 4가지 변환 유형에 대응되는 값을 1부터 4까지의 숫자로 줄 수 있다. val1과 val2 컬럼은 각 유형별로 규칙에 필요한 세부 데이터를 지정하기 위해 이용되며, 변환 유형별로 쓰임새가 달라진다.

위 예에서 “벌써”의 경우, 중립적인 의미를 가지는 부사에 해당하므로 type을 1로 정의하였으며, 이 경우 해당 단어는 S_{neut}에 추가된다. type이 1일 경우 val1과 val2에는 부가적인 값을 입력하지 않아도 된다. 이에 반해 type이 다른 값으로 정의될 경우, val1과 val2에 값이 필요할 수 있다. 예를 들어, “월요일은”은 명사를 포함하므로 조사를 분리하고 원형을 얻기 위해 type을 2로 정의한다. 그리고 분리될 조사를 지정하기 위해 val1 컬럼을 이용한다. 해당 예에서는 “은”을 val1에 지정함으로써, “월요일”이 원형으

로 분리될 수 있도록 정의한다. val1에 지정된 값은 카탈로그 업데이트 시 S_{post}에 추가된다. “해돋이”의 경우 type 2의 예외적인 형태로, “이”는 조사로 볼 수 없으며 단어에서 분리될 경우 오류가 발생한다. 따라서 “돋이”로 끝나는 단어의 경우 “이”가 분리되지 않도록 예외로 지정해야 하며, 이를 위해 val2에 “돋이”를 정의한다. val2에 지정된 값은 S_{poex}에 추가된다.

“귀여워”는 형용사의 활용형이므로 원형을 얻기 위해 type을 3으로 정의한다. 이 때 val1에는 해당 단어의 활용형 여부를 체크하기 위한 패턴을 정의하고, val2에는 해당 패턴을 만족할 경우 변환될 원형 단어를 정의한다. 정의된 값은 카탈로그 업데이트 시 [val1, val2]의 형태로 S_{conj}에 추가된다. val1에는 가급적 다양한 활용형 케이스를 포함할 수 있도록 정의하는 것이 좋다. 예를 들어, val1을 “귀여워”로 정의한다면, 단어가 “귀여워”로 시작되는 경우에 한해 “귀여운”으로 변환한다. 이에 반해, val1을 “귀여”로 정의하면 “귀여워” 뿐 아니라 “귀여워서”, “귀여워도” 등의 더욱 다양한 활용형을 포함시킬 수 있으며, 이 경우 규칙의 수를 줄이는 동시에 효율 향상을 기대할 수 있다. 마지막으로 “초콜렛”은 더욱 대중적으로 이용되는 유의어인 “초콜릿”으로 변환하기 위해 type 4로 정의한다. 이 경우 변환 대상 단어를 val1에 정의하며, 카탈로그 업데이트 시 [word, val1]의 형태로 S_{syn}에 추가된다.

엑셀 파일에 단어별 변환 규칙 정의가 완료되면, 이를 기반으로 카탈로그 업데이트 작업이 수행된다. 그리고 업데이트된 카탈로그 규칙을 기반으로 형태소 분석기가 입력 단어 집합 S에 대한 형태소 분석을 수행하고 최종 키워드 추출 결과를 출력한다.

IV. Experimental Results

이 장에서는 제안 방법에 대한 성능 및 정확도를 검증하기 위해, 온라인에서 수집된 실제 리뷰 데이터를 이용하여 수행된 실험 결과를 제시한다. 비교 대상과 관련하여, 이전 연구[6]에서 기존 형태소 분석기와의 성능 비교 결과를 제시한 바 있으므로, 본 논문에서는 이전 연구와 제안하는 방법의 성능을 비교하는 것으로 국한하였다.

1. Performance Test

제안 방법은 이전 연구의 키워드 추출 정확도를 보완하기 위한 목적으로 제시되었으므로, 비교에 있어 정확도가 어느 정도 개선되었는지, 그리고 제안 방법이 후처리의 형태로 적용될 때 어느 정도 추가 시간이 발생하는지, 여전히 실시간 처리 수준을 만족하는지 등의 관점에서 성능을 확인하고자 하였다.

실험을 위해 울산의 유명 스토어를 대상으로 네이버 검색 API를 통해 리뷰를 수집하였다. [표 3]은 울산의 각 지역구 별로 50개의 스토어를 대상으로 수집된 리뷰수와 형태소 분석 단계 적용 전/후의 키워드 오류 수를 비교한 실험 결과를 보여주고 있다.

Table 3. Error ratio before and after applying the proposed method

	남구	동구	북구	울주	중구
리뷰수	6,209	5,649	4,482	4,616	4,278
키워드수	1,242	927	816	464	782
적용전 오류수	125	117	125	63	54
비율	10.06%	12.62%	15.32%	13.58%	6.91%
적용후 오류수	4	2	8	6	2
비율	0.32%	0.22%	0.98%	1.29%	0.26%

형태소 분석 적용 전의 오류는 크게 두 가지로, 명사로부터 조사가 분리되지 않았거나, “있는”, “있어” 등 동사의 활용형에 대한 원형 추출이 제대로 되지 않은 경우가 많았다. 한편 형태소 분석 단계를 적용하여 이를 개선할 경우, 평균 10%에 이르던 오류율이 1% 이내로 개선되었다. 형태소 분석 적용 후에도 여전히 발견되고 있는 오류 케이스는 대부분 “삼산의”, “조민석의” 등과 같이 고유명사와 조사가 결합된 형태였다.

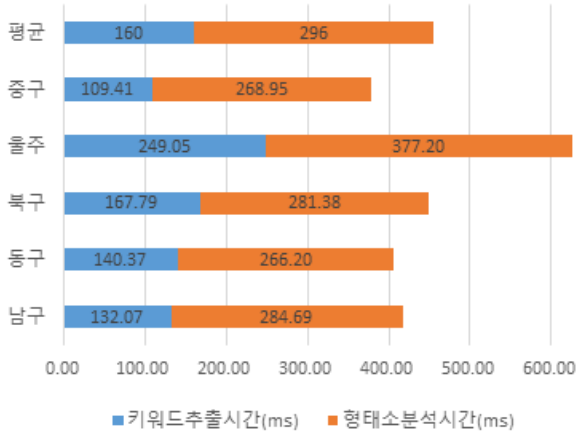

Fig. 3. Execution times of keyword extraction and the proposed morphological analysis (unit: milliseconds)

[그림 3]은 동일한 리뷰 집합에 대한 키워드 추출 시간과 형태소 분석 시간을 비교하였다. 비교의 편의를 위해

1,000개의 키워드를 처리하는데 걸리는 시간을 기준으로 값을 정규화하여 비교하였다. (1,000개의 키워드는 약 5,000개의 리뷰로부터 추출되는 양에 해당한다.) 실험 결과, 키워드 추출 시간에 비해 제안 방법의 형태소 분석 시간이 평균 54% 정도 더 차지하는 것으로 조사되었다. 그럼에도 불구하고, (5,000개의 리뷰를 대상으로) 둘을 합친 평균 처리 시간은 450ms 정도이며, 이는 제안 방법이 여전히 실시간 처리가 가능한 수준임을 보여준다.

2. Workload for Adding New Rules

앞서 언급한 바와 같이, 제안 방법에서는 입력 데이터가 주어질 때마다 형태소 분석 규칙을 수작업으로 추가해야 하며, 제안 방법의 효용성을 높이기 위해서는 해당 수작업에 들어가는 노력을 최소화시키는 것이 무엇보다도 중요하다. 이 절에서는 대량의 리뷰를 대상으로 분석 규칙에 소요되는 노력을 정량적으로 측정하기 위한 과정과 실험 결과에 대해 소개한다.

Table 4. Experimental data collected from 43,978 famous stores in Korea

	스토어수	리뷰수	토큰수	키워드 그룹수
울산	1,106	113,356	1,616,307	25,357
제주	1,642	362,347	5,257,494	91,217
부산	3,271	451,638	6,732,568	112,189
경남	3,039	397,454	5,977,230	101,777
강원	3,799	638,653	9,398,253	161,625
경북	3,571	450,114	6,517,249	109,973
대구	1,484	188,790	2,881,169	48,638
인천	1,594	246,623	3,649,798	62,553
서울	5,343	874,537	13,802,410	243,601
경기	6,517	875,603	13,494,367	224,171
충북	1,799	186,805	2,743,566	45,098
대전	1,313	120,558	1,875,952	30,486
충남	2,758	284,940	4,172,474	69,782
전북	2,483	295,159	4,411,005	74,649
광주	871	77,296	1,110,831	17,353
전남	3,388	399,024	5,869,071	101,504
총합	43,978	5,962,897	89,509,744	1,519,973

[표 4]는 전국 약 44,000여 개의 유명 스토어를 대상으로 네이버 검색 API를 통해 수집된 광역별 리뷰 수와 해당 리뷰로부터 얻어진 토큰 및 키워드 그룹의 개수를 보여준다. 토큰은 리뷰로부터 공백을 기준으로 분리하여 얻어진 단어 중, 특수문자 등을 제외한 단어에 해당한다. 키워드 그룹은 스토어별로 주어진 토큰으로부터 동일한 어근으로 시작되는 토큰들을 묶은 집합이다. 예를 들어 {"그라파", "그라파피자", "그라파피자리아"} 등을 포함한 단어 집합은 "그라파"로 시작하는 키워드 그룹으로 구성된다.

키워드 그룹이 생성되면 응집도 점수를 이용하여 그룹 내 단어 중 하나를 대표 키워드로 선정한다. 그리고 빈도수가 5이하인 키워드 그룹이나, 대표 키워드가 스토어 또는 지역 이름과 동일할 경우 해당 그룹은 키워드 추천에서 제외된다(자세한 내용은 이전 연구[6] 참조). 따라서 응집도 점수 적용 후 실제적으로 추천되는 키워드 수는 키워드 그룹 수에 비해 줄어들게 된다. 아래 표에서는 해당 값에 대해 "추천 키워드 수"로 표기하였으며, 이는 [그림 1]에서 |S|에 해당된다.

Table 5. Number of keywords reduced by applying the proposed interactive rule addition

	키워드 그룹수	추천 키워드수	신규 키워드수	신규키 비율
울산	25,357	3,036	1,546	50.9%
제주	91,217	4,149	1,455	35.1%
부산	112,189	8,377	2,234	26.7%
경남	101,777	8,257	2,362	28.6%
강원	161,625	10,063	2,626	26.1%
경북	109,973	8,867	2,141	24.1%
대구	48,638	4,443	945	21.3%
인천	62,553	5,138	1,174	22.8%
서울	243,601	15,539	2,999	19.3%
경기	224,171	20,609	3,531	17.1%
충북	45,098	4,389	777	17.7%
대전	30,486	3,241	559	17.2%
충남	69,782	6,752	1,201	17.8%
전북	74,649	5,419	962	17.8%
광주	17,353	2,399	430	17.9%
전남	101,504	8,110	1,565	19.3%
총합	1,519,973	118,788	26,507	22.3%

키워드 추출기로부터 S가 주어지면, 제안 방법은 내부 사전을 이용하여 신규 단어 집합인 R을 추출해 낸다. [표 5]는 S와 R의 크기를 비교하고 있으며, R의 크기에 대해 "신규 키워드 수"로 표기하고 있다. 위 실험 결과에서 볼 수 있듯이, 광역별로 실험이 진행될수록 R의 크기가 점차 감소하여, 규칙 추가에 필요한 수작업이 감소함을 알 수 있다. 예를 들어, 울산에서 시작할 때 R의 비율은 전체의 50.9%인데 반해, 실험 후반부에는 그 비율이 18%대로 줄어든 것을 확인할 수 있다. 즉 [그림 1]의 사전 구조를 통해 신규 키워드만을 확인할 수 있도록 지원함으로써, 키워드 확인에 필요한 노력을 약 80%까지 절감할 수 있음을 알 수 있다.

결과적으로, 전체 리뷰에 대한 형태소 분석 규칙 추가를 위해 확인해야 하는 키워드 수는 약 26,000여 개였으며,

추가된 변환 규칙 수는 [표 1]에서 보여진 바와 같이 1,700여 개로, 전체 키워드의 6.5% 정도만 변환 규칙을 가지는 것으로 조사되었다. 따라서 키워드 1개당 처리 시간을 5초 정도로 보수적으로 가정할 경우, 전체 키워드 처리에 약 36시간이 필요한 것으로 추정할 수 있다.

위 과정을 통해 초기 사전 및 분석규칙카탈로그가 완성되고 나면, 월별로 업데이트되는 신규 리뷰에 대한 키워드 확인 및 규칙 추가에 소요되는 노력은 대폭 단축된다. 아래 표는 동일한 스토어를 대상으로 최근 8월 업데이트된 리뷰로부터 얻어진 신규 키워드 수와 추가된 형태소 규칙 수를 보여준다.

Table 6. Number of new keywords and morpho logical analysis rules for the updated reviews collected in the latest August

	기존 키워드수	신규 키워드수	신규 규칙수	신규키 비율
울산	3,036	21	1	0.69%
제주	4,149	50	4	1.21%
부산	8,377	83	4	0.99%
경남	8,257	96	2	1.16%
강원	10,063	121	5	1.20%
경북	8,867	90	1	1.01%
대구	4,443	54	2	1.22%
인천	5,138	31	1	0.60%
서울	15,539	78	3	0.50%
경기	20,609	143	4	0.69%
충북	4,389	112	9	2.55%
대전	3,241	15	1	0.46%
충남	6,752	57	4	0.84%
전북	5,419	43	4	0.79%
광주	2,399	13	1	0.54%
전남	8,110	50	2	0.62%
총합	118,788	1,057	48	0.89%

실험 결과, 업데이트로부터 얻어진 신규 키워드 수는 1,057개로 전체의 0.89%에 해당하였으며, 이로부터 추출된 규칙 수는 48개로 신규 키워드 수의 4.5%를 차지하였다. 따라서 키워드 당 5초의 평균 처리 속도를 가정할 경우, 1.5시간 이내에 신규 키워드의 처리가 가능한 것으로 추정할 수 있다. 따라서 초기 사전 및 카탈로그 완성 후 업데이트된 리뷰를 처리하는데 드는 노력은 비교적 크지 않음을 알 수 있다.

V. Conclusion and Future Work

본 논문에서는 이전 연구로 수행되었던 응집도 점수 기반 키워드 추출의 정확도를 높이기 위한 보완책으로 대화형 형태소 분석기를 활용한 방법을 제시하였다. 제안 방법은 입력 데이터가 주어질 때마다 이를 처리하는데 필요한 형태소 분석 규칙을 점증적으로 추가할 수 있으며, 해당 구조를 통해 사전에 유지되는 단어와 분석 규칙의 수를 최소로 유지함으로써 실시간 처리가 가능하도록 구성하였다. 제안 방법의 검증을 위해 온라인에서 수집된 소셜 리뷰를 대상으로 실험을 수행하였으며, 제안 방법을 적용할 경우 기존 10% 오류율이 1% 이내로 개선됨을 확인하였다. 또한 5,000개의 리뷰 처리에 450ms 정도 소요되어 실시간 처리가 가능한 수준임을 확인하였다.

이에 반해, 제안 방법에서는 입력 데이터 처리에 필요한 변환 규칙을 수작업으로 추가해야 되는 문제점이 있었다. 따라서 해당 작업을 반자동화한 대화형 규칙 입력 시스템을 제공하여 문제를 해결하고자 하였다. 제안 방법의 검증을 위해 온라인에서 수집된 약 600만 개의 소셜 리뷰에 대한 실험을 수행하였으며, 그 결과 형태소 분석 규칙 추가와 사전 완성에 약 36시간 정도 걸리는 것으로 조사되었다. 또한 초기 사전 구성 후, 월별로 업데이트되는 신규 리뷰에 대한 처리 작업은 1.5시간 이내에 처리 가능한 것으로 파악되어, 업데이트가 진행될수록 규칙 추가에 드는 노력이 크게 줄어들게 됨을 알 수 있었다.

단, 제안 방법은 띄어쓰기가 잘 되지 않은 키워드를 처리하는 경우 여전히 오류가 발생될 수 있다. 예를 들어, 키워드가 “귀여운고양이”로 주어질 경우, 제안 방법은 3.2절에서 소개한 바와 같이 부분 패턴 매칭을 통해 “귀여운”을 결과로 출력하며, “고양이” 정보는 사라지게 된다. 따라서 관련 오류를 보완하기 위한 추가 연구를 지속적으로 수행할 예정이다.

ACKNOWLEDGEMENT

This research was supported by the Sahmyook University Research Fund in 2020.

REFERENCES

[1] H. G. Kim, "Developing a Big Data Analysis Platform for Small and Medium-Sized Enterprises," Journal of the Korea Society of

Computer and Information, Vol. 25, No. 8, Aug. 2020.
[2] W. L. Kang, H. G. Kim, and Y, J. Lee, "Reducing IO Cost in OLAP Query Processing with MapReduce," IEICE Trans. Inf. & Syst, Vol. E98-D, No. 2, pp. 444-447, Feb. 2015.
[3] Naver Open API, https://developers.naver.com/docs/common/ open apiguide/
[4] Google Developer API, https://developers.google.com/
[5] Kokoma, http://kkma.snu.ac.kr/documents/index.jsp
[6] H. G. Kim, "Efficient Keyword Extraction from Social Big Data Based on Cohesion Scoring," Journal of the Korea Society of Computer and Information, Vol. 25, No. 10, Oct. 2020.
[7] H. Lim, B. Yoon, and H. Lim, "An Efficient Korean Morphological Analyzer using Exclusive Information," Journal of KIISE, Vol. 22, No. 6, pp. 957-964, 1995.
[8] Y. Kim, M. Park, J. Choi, and H. Kwon, "Improvement of Analysis Speed in Korean Morphologlcal Analyzer Using Ameliorated Dictionary," Proc. of the 11th Hangul and Korean Information Processing, pp. 479-483, 1999.
[9] S. H. Yang and Y. S. Kim, "A High-Speed Korean Morphological Analysis Method based on Pre-Analyzed Partial Words," Journal of KIISE, Vol. 27, No. 3, pp. 290-301, 2000.
[10] H. G. Seo and H. W. Park, "Design and Implementation of Potential Advertisement Keyword Extraction System Using SNS," Journal of the Korea Convergence Society, Vol. 9, No. 7, pp. 14-24, 2018.
[11] O. J. Lee, S. B. Park, D. Chung, and E. S. You, "Movie Box-Office Analysis Using Social Big Data," Journal of the Korea Contents Society, Vol. 14, No. 10, pp. 527-538, 2014.
[12] C. Lee, D. Choi, S. Kim, and J. Kang, "Classification and Analysis of Emotion in Korean Microblog Texts," Journal of KIISE, Vol. 40, No. 3, pp. 159-167, Jun. 2013.
[13] J. Y. Chang, "A Sentiment Analysis Algorithm for Automatic Product Reviews Classification in Online Shop ping Mall," Vol. 14, No. 4, pp. 19-32, 2009.
[14] Hannanum, http://semanticweb.kaist.ac.kr/hannanum/index.html
[15] Z. Jin and K Tanaka-Ishii, "Unsupervised Segmentatino of Chinese Text by Use of Branching Entropy," The Journal of Korea Navigation Institute, pp. 428-435, Jul. 2006.
[16] H. J. Kim and S. J. Cho, "Cleansing Noisy Text Using Corpus Extraction and String Match," MS. Thesis, Seoul National University, 2013.
[17] Cohesion Score, https://lovit.github.io/nlp/2018/04/09/cohesion _ltokenizer/
[18] Soynlp, https://github.com/lovit/soynlp
[19] E. Kim, "The Unsupervised Learning-based Language Modeling of Word Comprehension in Korean," Journal of the Korea Society of Computer and Information, Vol. 24, No. 11, pp. 41-49, Nov. 2019.

Authors

Yang Woo Yu received the B.S. M.S. and Ph.D degrees in Computer Science from University of Ulsan, Korea, in 1995, 1997 and 2005, respectively Dr. Yu joined the Department of the Digital Contents Design at Ulsan College, Ulsan, Korea, in 2000. He is currently an Associate Professor in the Department of the Digital Contents Design at Ulsan College, He is interested in VR programming, web programming and big data processing.

Hyeon Gyu Kim received the B.S. and M.S. degrees in Computer Science from University of Ulsan, and Ph.D. degree in Computer Science from Korea Advanced Institute of Science and Technology, Korea, in 1997, 2000 and 2010, respectively. Dr. Kim joined the faculty of the Division of Computer Science and Engineering at Sahmyook University, Seoul, Korea, in 2012. He is currently an Associate Professor in the Division of Computer Science and Engineering, Sahmyook University. He is interested in big data processing, data stream processing, and mobile computing.