https://doi.org/10.5392/JKCA.2024.24.07.082

레벤슈타인 편집 거리 알고리즘을 이용한 법률 용어 오탈자 자동 교정 시스템 제안

Auto-correction with Levenshtein (Edit) Distance Algorithm to Spell Error in Legal Terms

공성호*, 심진우, 현민호*, 문성민*****

아주대학교 사학과*, 아주대학교 문화콘텐츠학과**, 아주대학교 인문과학연구소***

Sung-Ho Gong(sam2638@ajou.ac.kr)*, Jin-Woo Shim(deunsol1227@ajou.ac.kr), Min-Ho Hyeon(gusalsgh312@ajou.ac.kr)*, Seongmin Mun(stat34@ajou.ac.kr)*****

요약

법제처에서는 법률 용어 순화 작업을 진행하고 있지만, 실상은 어려운 법률 용어가 계속 사용되고 있어 사용자가 법률 용어를 이해하거나 사용하기 위해서는 높은 법학지식 수준이 요구된다. 따라서 본 연구는 어려운 법률 용어에 대한 사용자들의 접근성을 높이고자 Levenshtein (Edit) Distance 알고리즘을 이용하여 잘못 사용된 법률 용어를 자동으로 교정해주는 시스템을 고안하여 제시한다. 연구를 위한 데이터로는 한국법제연구원(KLRI)의 법령 용어 사전, 대한민국 법원에서 제공하는 880개 판례에서 사용된 주요 단어들, 그리고 국립국어원에서 제공되는 약 100만 개의 한글 단어 사전이 연구의 데이터로 사용되었고 연구의 방법으로는 한국어의 음소를 기반으로 입력된 단어와 실제 법률 용어 사이의 편집 거리를 계산하는 Levenshtein (Edit) Distance 알고리즘이 사용되었다. 연구의 결과, 제안된 시스템은 잘못 사용된 법률 용어를 교정하는 데 있어 평균 96%의 정확도를 보여주었다. 또한, 실제 36명의 사용자를 대상으로 진행한 설문조사에서 정확성, 용이성, 정보성, 사용성에 있어 모두 긍정적인 평가 결과를 보여주었다.

■ 중심어 : | 자동 교정 | 법률 용어 | 오탈자 교정 | 편집 거리 | 자연어 처리 |

Abstract

The Ministry of Government Legislation in Korea is currently undergoing a simplification process of legal terminology. However, in reality, complex legal terms are still being used, and these terms demand a high level of legal expertise for users to comprehend or utilize them. Therefore, this study proposes a system designed to enhance the accessibility of users to challenging legal terms by automatically correcting misused legal terms using the Levenshtein Edit Distance algorithm. For the data, we utilized the legal terminology dictionary from the Korea Legislation Research Institute (KLRI), key words extracted from 880 cases provided by Court of Korea, and approximately one million Korean words from the National Institute of Korean Language. We then employed the Levenshtein Edit Distance algorithm, which calculates the editing distance between input words based on Korean phonetics and legal terms. The results of the study demonstrated an average accuracy of 96% in correcting misused legal terms. Furthermore, a survey conducted with 36 participants yielded positive evaluations in terms of accuracy, ease of use, informativeness, and usability of the proposed system.

■ keyword : | Auto Correction | Legal Terms | Spell Error Correction | Edit Distance | Natural Language Processing |

* 본 연구는 한국연구재단의 인문사회분야 인문사회연구소지원사업의 지원(NRF-2022S1A5C2A02090368)을 받아 수행되었습니다.

접수일자 : 2024년 04월 03일
심사완료일 : 2024년 05월 03일
수정일자 : 2024년 05월 03일
교신저자 : 문성민, e-mail : stat34@ajou.ac.kr

I. 서 론

우리나라의 법률 용어는 한자문화권인 우리나라의 특성과 일제강점기를 겪은 역사를 통해 자연스럽고 알기 쉬운 우리말 표현이 아닌 한자어, 일본식 용어, 전문용어, 어색한 용어 등이 섞여 실제로 그 의미를 이해하기 힘든 용어들이 포함되어 있다. 따라서 법제처는 2006년부터 법령을 알기 쉽게 정리하여 국민 눈높이에 맞는 친근한 법령 만들기를 목표로 '알기 쉬운 법령 만들기(알법)' 사업을 꾸준하게 진행해왔다[1]. 해당 사업은 단기간으로 끝나지 않고 현재도 진행되고 있으며 국민들의 법령 사용과 해석에 대한 어려움을 낮추는데 기여하고 있다. 법제처와 별도로 법원도 이에 대한 노력을 기울이고 있다. 법원은 판결서 작성이나 법원 내외의 문서, 용어 사용의 문제를 인식하고 해결하고자 노력하고 있으며, 1997년 12월 법원 맞춤법 자료를 발간한 후, 2010년 읽기 쉬운 판결서 작성 핸드북 발간, 2013년엔 법원 맞춤법 자료집 발간하는 등의 노력을 기울이고 있다[2-4].

한국 정부 기관의 직접적인 노력 외에도 법률 용어에 대한 사용자들의 접근성을 높이기 위한 다양한 사전 연구들도 진행되어왔다. 예를 들어, [5]는 일반 사용자들이 사용하는 생활 용어를 기반으로 이에 적합한 법률 용어를 추출하고 법률정보를 직접 찾을 수 있도록 하는 시스템을 제안하였다. 연구의 방법으로 해당 연구는 블로그의 태그 정보를 데이터로 사용하였다. 그 뒤, 태그 정보에 대한 군집화(clustering)를 수행하고 군집화된 태그 정보를 기반으로 입력된 생활 용어와 대응하는 법률 용어를 추출하고 법률정보를 검색할 수 있도록 도움을 주는 시스템을 제안하였다[5].

다음으로 [6]은 일반 사용자들이 사용한 이용대금을 직접 법률문서로 작성하는 데 도움을 주고자 이를 자동 작성해주는 서비스 제안하고 자동 작성된 문서가 법률 사무인지를 검증하는 연구를 진행하였다. 사용자가 변호사를 통해 자신이 사용한 이용대금에 대한 법률문서를 작성하기 위해서는 금전적으로 큰 부담이 되지만 해당 연구 방법이 변호사법상 문제가 될 수 있는 법률 사무에 해당하지 않는다면 제안된 연구의 서비스를 통해 많은 사용자에게 도움이 될 수 있다[6].

앞서 언급된 정부의 노력과 사전 연구들의 방법에도 불구하고 아직도 일반 사용자가 법률 용어를 이해하거나 사용하기 위해서는 높은 법학지식 수준이 요구된다는 문제점이 있다. 따라서 본 연구는 어려운 법률 용어에 대한 사용자들의 접근성을 높이고자 Levenshtein (Edit) Distance 알고리즘을 이용하여 잘못 사용된 법률 용어를 자동으로 교정해주는 시스템을 고안하여 제시하고자 한다.

Levenshtein (Edit) Distance(이하 LD)알고리즘은 두 문자열의 유사도를 측정하는 방법으로 오탈자 교정을 위한 알고리즘으로 많이 사용된다. 예를 들어, [7]는 도서관에서 사용되는 도서 검색 시스템의 성능을 향상시키고 사용자에게 더 나은 검색 환경을 제공하고 LD 알고리즘을 사용하여 자동완성 및 철자 검사 알고리즘을 시스템에 도입하였다. 다음으로 [8]은 영화에 대한 평가 리뷰에서 감성을 분류하는 연구를 진행하였는데, 데이터를 정제하는 과정에서 더 정확한 텍스트를 수집하여 사용하고자 LD 알고리즘을 사용하였다. 이 외에도 [9]의 연구에서는 각 어휘의 실제 사용 빈도에 따라 가중치를 부여하고 이를 거릿값에 적용한 개선된 LD 알고리즘을 제안하였으며, 이를 기반으로 음성인식 기능의 정확도를 향상시키는 연구를 진행하였다.

선행 연구들에서 보고된 LD알고리즘의 오탈자 교정 성능을 고려하여, 본 연구는 LD 알고리즘을 사용하여 법률 용어의 오탈자를 자동으로 교정하는 시스템을 제안하는 것을 목적으로 연구를 수행하였다. 본 연구의 목적을 위한 연구 절차는 다음과 같다. 첫째, 본 연구는 연구 수행을 위한 데이터로써 국립국어원의 우리말샘 데이터, 법령용어 검색 서비스, 대한민국 법원 종합 법률 정보 사이트 등을 말뭉치 데이터로써 수집한다. 둘째, 수집된 말뭉치를 사용자 사전으로 전환하여 LD 알고리즘에 적용한다. 추가적으로 LD 알고리즘을 한국어에 적합하게 적용하여 작동할 수 있도록 형태소 분석, 음절, 음소 분리 등 추가적인 개발을 진행한다. 셋째, 마지막으로 개발된 시스템의 성능 평가를 위해 일반 사용자 집단을 모집하고 시스템에 대한 성능 평가를 진행한다.

II. 사전 연구

1. 법률 분야에서 사용자의 편의를 위한 연구

법률 분야에서 사용자의 편의를 위한 연구는 계속되어 왔다. 법제처에서 진행하는 '알기 쉬운 법령 만들기'를 통해 일본어식 용어나 한자어가 많은 법령의 한글화, 어려운 법령용어의 순화, 간결화와 명확화 등 법률의 접근성을 높이고 있다[1].

[5]는 기존 법률 용어의 어려움을 극복하고, 일반 사용자도 법률 정보를 쉽게 검색할 수 있도록 도와주는 방법론을 제안하며, 생활 용어와 법률 용어 간의 대응관계를 탐색하여 검색에 활용할 수 있는 방법을 제시한다. 이를 위해 한국 최대 포털사이트 네이버의 집단 지성을 활용하여 블로그와 티스토리에 작성되는 태그 정보를 수집하는 방안을 제시한다. 이런 태그 정보를 수집하기 위해서는 생활 용어를 검색하였을 때 나오는 블로그의 태그 정보를 수집하고, 태그의 용어 구분을 위하여 법률 검색 사이트에 태그 정보를 검색한다. 검색 결과가 나올 경우 법률 용어, 나오지 않을 경우는 생활 용어로 정의한다. 수집된 태그 데이터 중 밀접하게 관련된 태그를 추출하기 위해 검색어와 동시 출현 빈도가 3 이상인 태그만을 추출한다. 그리고 수집된 태그를 벡터화한 후 k-mean 클러스터링(군집화)을 수행한다. k-means 클러스터링을 통해 생활 용어에 대응되는 법률 용어의 집합을 찾을 수 있고, 생활 용어에 대응되는 하나의 법률 용어를 찾기 위해 Lift-value를 사용하여 생활 용어와 법률 용어의 대응 관계를 판단하며, Lift-Value 값이 큰 법률 용어를 선택하는 방법을 사용한다. 이를 통해 생활 용어에 대응하는 하나의 법률 용어를 찾고 SKOS를 활용하여 온톨로지를 구축한다. 해당 알고리즘의 성능을 평가하기 위해 자주 사용하는 40개 정도의 생활 용어를 선정하여 알고리즘에 적용한다. 대부분의 생활 용어가 적합한 법률 용어와 매칭되었으나 "복덕방", "부가세", "시골땅"과 같은 애매한 단어들은 적합한 법률 용어와 매칭되지 않았다. 해당 논문은 블로그 사용자들이 태그를 자세하게 기재하지 않는 한계점이 원인이라고 하였지만, 이는 검색에 있어서 태그가 점점 중요해지는 추세에 따라 해결될 수 있다고 본다. 이 연구는 블로그의 태그 정보를 활용하여 태그 클러스터링을 통해 생활 용어와 법률 용어 간의 대응관계를 자동으로 탐색하는 방법을 제시했다. 이를 통해 사용자들은 법률에 대한 정확한 지식이 없어도 생활 용어를 질의어로 사용하여 법률 정보를 얻을 수 있는 기반을 마련했다. 그러나 현재의 문제점은 시소러스(용어집)의 법률 용어를 보완해야 한다는 점이다. 클러스터링을 통해 얻은 법률 용어가 시소러스에 없다면 생활 용어와의 연결이 불완전할 수 있다. 또한, 연구는 생활 용어 선택에서의 수동성을 줄이기 위해 자동으로 생활 용어와 이에 해당하는 법률 용어를 탐색하여 온톨로지를 구축하는 것을 향후 과제로 삼고 있다[5].

2. Levenshtein Distance

LD 알고리즘은 1965년 Vladimir Levenshtein에 의해 고안되었고, 두 개의 문자열의 유사도를 계산하는데 자주 사용된다. LD는 한 단어를 다른 단어로 변경하는데 필요한 편집의 최소 개수이고 편집의 종류는 삽입(Insertion), 삭제(Deletion) 그리고 대체(Substitution)가 있다. LD 알고리즘은 큰 문제를 작은 문제들로 나누어 풀어나감으로써 해결하는 동적 계획법(Dynamic Programming)으로 계산된다.

[7]는 도서관 도서 검색 시스템에 LD알고리즘을 사용하여 자동완성 및 철자 검사를 적용해 사용자에게 더 나은 검색 환경을 제공하고자 하였다. 해당 시스템의 자동완성 기능의 목적은 시스템 사용자가 최소한의 입력으로 적절한 책 추천을 받을 수 있도록 함이고, 철자 검사 기능은 사용자의 입력에 오류가 있을 때 자동 교정과 함께 책 추천을 제공받는 것이다. 해당 연구에 사용된 데이터는 UNNES 도서관의 다양한 분야의 도서 데이터로서 2000년부터 2017년까지의 데이터가 포함되었고, 총 1155권의 도서 제목과 설명이 수집되었다. 해당 연구는 도서 시스템에 자동완성과 철자 검사 기능을 위하여 LD 알고리즘을 사용하였다. 자동완성 기능은 사용자가 시스템에 도서 이름을 입력하고 도서관의 데이터베이스와 LD를 비교하여 일치 여부를 확인한 후 사용자에게 자동 완성을 제공한다. 그리고 철자 검사 기능은 사용자가 잘못된 도서 이름을 입력하였을 EO 자동완성과 마찬가지로 데이터베이스와 비교하여 LD가 최소인 도서를 추천한다. 해당 연구에서는 구현된

시스템의 성능을 확인하기 위해 100개의 테스트 데이터와 1055개의 훈련 데이터를 사용하여 검증을 수행하였다. 테스트 결과 철자 검사의 정확도는 86%로 나타났으며 검색 정확도를 향상해 사용자에게 더 좋은 검색 환경을 제공할 수 있음을 보여주었다[7].

[8]은 영화평의 감성 분류를 위해 LD를 사용하는 방법을 제안하였다. 이 연구는 철자 오류에도 영향을 적게 받는 감성 분류 방법을 제시하고 있다. 8,182개의 영화평 데이터를 사용했고 천만 명 이상의 관객을 가진 영화에서 수집되었으며, 긍정과 부정이 확실한 2,385개의 데이터를 선별하여 사용하였다. 총 수집된 감성 어휘는 778개다. [8]에서는 LD를 이용하여 BOW(bag of words)를 생성하고 다항 나이브 베이즈 분류기에 적용하여 학습을 진행하였다. 실험 결과 LD가 3일 때 가장 높은 정확도를 보였다. 실험 결과를 통해 [8]에서 제안하는 방법은 철자 오류 및 띄어쓰기 오류의 영향을 감소시키고 감성 분류 성능을 높일 수 있음을 보였다.

LD의 응용 분야로, [10]은 LD 알고리즘을 휴대폰 카메라를 이용한 간판 영상에서 한글 인식에 적용하였다. 제안된 wDLD(weighted Disassemble Levenshtein Distance)를 통해 인식 후보의 음소를 분리하고 연산에 가중치를 적용하여 정확한 상호명을 검출하는 방법을 제시하였다. 실험 결과에 따르면 해당 방법은 기존 방법에 비해 평균 29.85%, 6%의 인식률 향상을 보였다[10]. 이러한 다양한 연구들은 LD알고리즘이 철자 오류를 교정하거나 문자열을 비교하는데에 효과가 있음을 시사하고 있다.

[9]는 음성 인식에 주로 사용되는 HMM(Hidden Markov Model) 모델에 기존 LD 알고리즘의 단점인 어휘의 우선 순위가 없다는 점을 보완하여 음성 인식 시스템의 성능을 향상하고자 한다. 음소 간 변화에 대한 가중치를 다르게 적용하여 개선하였고, 음소열을 사용하므로 어휘가 증가하여도 인식률이 향상된다. 또한 어휘들의 우선 순위를 매기기 위해 가중치를 부여한다. 사용 빈도에 따라 어휘를 탐색하도록 처리하여 인식률과 인식 시간을 향상하도록 하였다. 제안한 시스템의 성능을 평가하기 위해 기존 방식과의 비교 실험을 수행하였다. 실험은 화자 종속형과 화자 독립형으로 분류되었으며, 결과적으로 실내 환경에서 어휘 종속 및 독립 인식률은 각각 97.81%, 96.91%로 Viterbi 알고리즘 대비 2.3% 향상되었다. 또한, 실외 환경에서는 어휘 종속 및 독립 인식률이 각각 91.11%, 90.01%로 Viterbi 알고리즘 대비 1.1% 향상되었다. 이를 통해 음소열을 사용하고 어휘에 가중치를 부여한 개선된 Levenshtein distance 알고리즘이 효과적임이 입증되었다.

[33]은 영문교정에 있어서 일반적인 영문이 아닌 전문적이고 학술적인 영문을 자동으로 교정하여 제공하는 방법론을 제시한다. 해당 연구는 두 단계로 이루어져 있다. 첫 번째는 오탈자를 교정하는 단계. 두 번째는 문장의 전달력을 개선하는 단계이다. 오탈자를 교정하기 위해서 우선 수집한 학술 용어 사전을 기반으로 오탈자를 감지하고 교정 후보를 생성한다. 다음으로 에러 모델을 사용하여 교정 후보 중 적합한 용어를 선정한다. 마지막으로 n-gram 언어 모델을 활용하여 최종적으로 선정된 용어를 사용자에게 제공한다. 문장의 전달력을 개선하는 단계에서는 딥러닝 기법인 양방향 순환 신경망(Bidirectional Recurrent Neural Network, BRNN) 기반의 자동 사후 교정(Automatic Post-Editing, APE) 모델을 제안한다. 해당 연구는 다양한 분야의 학술 용어를 위해 교정 전문가가 교정한 학술 영어 논문으로부터 329,374개의 학술 용어를 수집하였고, 생물 분야 학술 용어를 위해 Cspell 단어사전과 의학 전문 사이트 KMLE의 전문용어를 수집하였다. 오탈자 교정의 품질을 높이기 위해 3.7×10^8개의 웹페이지로부터 오탈자 데이터를 수집하였다. 제안된 모델은 영어 원문 X를 받는 인코더와 교정문 Y를 출력하는 디코더로 구성되어 있으며, 원문과 교정문 쌍으로 문장 간의 차이를 학습하도록 설계되었다. 또한 문장 길이에 구애받지 않고 교정이 가능하다. 해당 연구는 제안 방법론의 성능을 평가하기 위해 교정 전문가의 교정이 완료된 영문 데이터를 이용하여 단일 언어 사후 교정의 성능을 신경망 기계번역(Neural Machine Translation, NMT) 기반의 APE 시스템과 정량적 비교와 정성적 비교 두 가지 방법을 통해 실험을 진행하였다. 정량적 비교 실험에서 여섯 개의 분야 중 다섯 개 분야에서 높은 GLEU(Global Language Understanding Evaluation) 값을 보였고, 정성적 비교 실험에서 [33]의 방법론은 비교 방법론이 감지하지 못한 '기관지확장제(bronchodilator)'

라는 의학 용어의 오탈자를 감지하여 교정하였고, 문장 전달력 또한 개선된 것을 확인할 수 있었다.

III. 연구 방법

1. 데이터 수집

본 장에서는 데이터를 수집한 웹 사이트와 그 데이터를 정제하는 방법에 대해 설명한다.

사전 연구들에서 보고된 결과에 따르면 LD알고리즘을 기반으로 한 오탈자 교정 시스템의 정확도는 사전에 구축된 데이터의 양에 좌우되기 때문에 본 연구에서는 최대한 많은 데이터를 수집하여 연구에 사용하였다. 우선 국립국어원에서 2010년부터 시작한 '개방형 한국어 지식 대사전 구축' 사업의 대표 사전인 '우리말샘'을 이용하였다[그림 1].

"channel": {
"total": 50000,
"title": "사전 검색",
"description": "사전 검색 결과",
"item": [
{
"wordinfo": {
"word_unit": "어휘",
"origin": "< German",
"word": "게르만",
"original_language_info": [
{
"original_language": "←Geruman",
"language_type": "안 밝힘"
}
],
"word_type": "외래어"
},
"group_order": 1,
"group_code": 46558,
"link": "http://opendict.korean.go.kr/dictionary/view?sense_no=413165",
"target_code": 413165,
"senseinfo": {
"definition": "게르만 어파에 속한 언어를 쓰는 민족. 백색 인종으로 키가 크고 금발이며 눈이 푸
"cat_info": [
{
"cat": "고유명 일반"
}
],
"relation_info": [
{
"link": "http://opendict.korean.go.kr/dictionary/view?sense_no=523779",

그림 1. 우리말샘 JSON 파일 형식

국립국어원에서는 2016년에 "우리말샘" 사전을 공개하였고, 약 100만 개의 단어가 제공된다. 해당 사전은 JSON(Java Script Object Notation) 형태로 제공되고 있으며 파이썬(Python) 프로그래밍 언어(이하 파이썬)를 이용하여 데이터 정제 작업을 진행하였다. JSON 형태의 데이터를 파이썬의 딕셔너리(Dictionary) 자료형으로 변환하고(①) 단어의 기본형인 "word"의 값을 수집한 후(②) 텍스트 파일에 저장하였다(③). 또한 정부출연연구기관인 한국법제연구원(KLRI)의 법령용어검색서비스를 사용하였다[그림 2]. 해당 웹 사이트에 저장된 6,178개의 법률 용어를 수집하기 위해 파이썬의 BeautifulSoup[1]와 Selenium[2] 라이브러리를 이용하였다. 각 페이지에 있는 용어를 수집하고 다음 페이지로 넘기는 작업을 반복하여 실행하였고, 총 6,178개의 용어를 텍스트 파일에 저장하였다.

그림 2. 한국법제연구원(KLRI)의 법령용어검색서비스

마지막으로 실제 판례에서 사용하는 용어를 수집하기 위해 대법원 종합법률정보에 공개된 "화제의 판례" 880개를 활용하였다. 법령용어검색서비스와 같이 파이썬의 BeautifulSoup와 Selenium을 이용하여 판례를 수집하였고, Kiwi(Korean Intelligent Word Identifier) 지능형 한국어 형태소 분석기를 이용하여 내용어(내용어는 문장 안에서 명확한 어휘적 의미를 가지고 있다)인 명사, 형용사, 동사, 부사를 따로 추출해 수집하였다. 수집한 용어는 하나의 텍스트 파일에 저장하였고 중복된 용어를 제거해 최종적으로 1,013,644개의 용어가 저장된 14.8MB의 텍스트 파일 데이터를 확보하였다.

2. Levenshtein Distance

본 장에서는 Levenshtein Distance알고리즘을 사용하여 주어진 두 단어 사이의 편집거리를 계산하는 방법을 소개한다. LD는 한 단어를 다른 단어로 변경하는데 필요한 편집의 최소 개수를 의미하며 삽입(Insertion), 삭제(Deletion), 대체(Substitution)의 과정으로 두 단어의 편집거리를 계산하는 방법은 [그림 3]과 같다.

1 웹 사이트의 html코드를 추출하기 쉽게 해준다.
2 코드를 이용하여 웹 브라우저를 조작해준다.

```
① matrix = []
   for i in range(len(str1)+1):
      row = []
      for j in range(len(str2)+1):
         row.append(0)
      matrix.append(row)
② for i in range(1, len(str1)+1):
      matrix[i][0] = i
   for j in range(1, len(str2)+1):
      matrix[0][j] = j
③ for i in range(1, len(str1)+1):
      for j in range(1, len(str2)+1):
         if str1[i-1] == str2[j-1]:
            matrix[i][j] = matrix[i-1][j-1]
         else:
            matrix[i][j] = min(matrix[i-1][j], matrix[i][j-1], matrix[i-1][j-1]) + 1
```

그림 3. Levenshtein 알고리즘

[그림 3]의 편집거리 계산 방법을 예시를 통해 살펴보면 다음과 같다. [표 1]에서 “apple”과 “apricot’의 LD는 가로, 세로 길이가 “apple”과 “apricot”보다 1씩 긴 6X8의 영행렬을 생성한다(①). 행렬의 첫 행과 첫 열을 각 단어의 길이에 따라 1부터 끝까지 채운다(②). 편집이 행해지면 행과 열의 값은 왼쪽, 위, 왼쪽 대각선 위 3개의 값 중 가장 작은 값에 1을 더한 값으로 채워진다. “apple”를 “aprle” p를 r로 대체한다(③). “aprle"에서 “aprie”로 l을i로 대체한다(④). “aprie”를 “apric”로 e를 c로 대체한다(⑤). “apric”를 “apricot”로 o를 추가한다(⑥). “apricot”를 “apricot”으로 t를 추가한다(⑦). 5번의 편집이 이루어졌고, 행렬 끝의 숫자가 5가 나옴으로써 두 단어 사이의 LD는 5가 된다.

표 1. LD알고리즘 영어 단어 예시

		a	p	r	i	c	o	t
	0	1	2	3	4	5	6	7
a	1	0	1	2	3	4	5	6
p	2	1	0	1	2	3	4	5
p	3	2	1	1	2	3	4	5
l	4	3	2	2	2	3	4	5
e	5	4	3	3	3	3	4	5

3. 음소 분리

본 장에서는 법률 용어를 음소 단위로 분리하여 LD 알고리즘을 적용하는 것을 소개한다. 한글 단어를 LD 알고리즘에 적용하면 정확도가 많이 떨어진다. 예시로 “바나나”, “바람”, “바구니” 세 단어의 LD는 모두 2로 교정에 어려움이 있다. 이에 한국어는 음소단위로 구성되어 있다는 특성을 이용하였다. 파이썬의 python-jamo 라이브러리를 이용하여 음소단위인 자음과 모음을 분리하고 진행하였다. [표 2]에서 법률 용어인 “흡수합병”과 오탈자 “흖ㅜㅎ밥병”를 python-jamo를 이용하여 자음, 모음 단위로 분리한다(①). 분리된 두 문자열의 길이가 11로 12X12의 영행렬을 생성한다(②). 첫 행과 첫 열이 각 0부터 11까지 채워진다(③). “ㅎㅡㅃㅜㅎㅂㅏㅂㅂㅕㅇ”가 “ㅎㅡㅂㅜㅎㅂㅏㅂㅂㅕㅇ”으로 ㅃ이 ㅂ으로 대체된다(④). “ㅎㅡㅂㅜㅎㅂㅏㅂㅂㅕㅇ”가 “ㅎㅡㅂㅅㅜㅎㅂㅏㅂㅂㅕㅇ”으로 ㅅ이 추가된다(⑤). “ㅎㅡㅂㅅㅜㅎㅂㅏㅂㅂㅕㅇ”가 “ㅎㅡㅂㅅㅜㅎㅏㅂㅂㅕㅇ”으로 ㅂ이 삭제된다(⑥). 3번의 편집이 이루어졌고, 행렬 끝의 숫자가 3이 나옴으로써 LD는 3이다.

표 2. LD알고리즘 한글 단어 예시

		ㅎ	ㅡ	ㅃ	ㅜ	ㅎ	ㅂ	ㅏ	ㅂ	ㅂ	ㅕ	ㅇ
	0	1	2	3	4	5	6	7	8	9	10	11
ㅎ	1	0	1	2	3	4	5	6	7	8	9	10
ㅡ	2	1	0	1	2	3	4	5	6	7	8	9
ㅂ	3	2	1	1	2	3	3	4	5	6	7	8
ㅅ	4	3	2	2	2	3	4	4	4	6	7	8
ㅜ	5	4	3	3	2	3	4	5	4	6	7	8
ㅎ	6	5	4	4	3	2	3	4	5	6	7	8
ㅏ	7	6	5	5	4	3	3	3	4	5	6	7
ㅂ	8	7	6	6	5	4	3	4	3	4	5	6
ㅂ	9	8	7	7	6	5	4	4	4	3	4	5
ㅕ	10	9	8	8	7	6	5	5	5	4	3	4
ㅇ	11	10	9	9	8	7	6	6	6	5	4	3

4. 자동 교정 절차

본 장에서는 수집한 데이터 용어 사전과 음소 분리를 적용한 LD알고리즘을 설명한다. 본 연구에서 제안하는 시스템의 작업 절차는 [그림 4]와 같다. 우선, 사용자가 입력한 법률 용어가 시스템에 입력되고(①) 용어 사전에 입력된 법률 용어가 존재하는지 확인한다(②). 만약 사전에 입력된 용어가 존재하지 않는다면 오타라고 판단하여 사용자에게 입력받은 용어와 용어 사전에 있는 용어들 사이의 편집거리(LD)를 계산한다(③). 키(Key)가 용어, 값(value)이 LD로 이루어진 파이썬의 딕셔너리(dictionary) 자료형을 생성하고(④) 산출된 LD값을 기준으로 오름차순 정렬한다(⑤). 그리고 딕셔너리의 0번째 인덱스(index)에 위치한 용어를 사용자에게 출력한다(⑥). 그리고 만약 ②의 과정에서 사용자에게 용어를 입력 받을 때 용어 사전에 입력된 법률 용어가 존재한다면 오타가 아니라고 판단하여 입력받은 용어 그대로 출력하도록 한다.

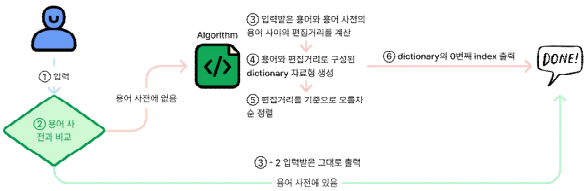

그림 4. 시스템 플로우차트

[표 3]은 본 연구에서 제안하는 법률 용어 교정 시스템을 통해 입력된 용어(교정 전)를 교정한 결과(교정 후)를 보여주고 있다.

표 3. 시스템의 입력과 결과

교정 전	교정 후
흡수하ㅂ병	흡수합병
좨수관계	죄수관계
근노계약	근로계약
가사비옹사건	가사비송사건
과새전적부심사	과세전적부심사

IV. Evaluation

1. 시스템 정확도 테스트

본 장에서는 제안된 LD 알고리즘의 정확도 테스트 결과와 용어 인식 경향에 대해 설명한다. 정확도 테스트를 위해 본 연구는 파이썬의 BeautifulSoup과 Selenium 라이브러리를 사용하여 대법원의 "대한민국 법령 종합 법률 정보" 웹 사이트에서 제공되는 판결문을 문자열 형태로 수집하여 테스트에 사용하였다. 시스템 정확도 평가에 수집된 판결문을 사용하기 위해 우선 텍스트 정제 과정을 진행하였다. 텍스트 정제 과정은 다음과 같다. 우선 Kiwi 라이브러리를 이용하여 수집된 판결문을 문장 단위로 분리하였다. 그 뒤 형태소 분석을 통해 문장에 존재하는 문맥어(동사, 형용사, 명사, 부사)들만을 추출하여 시스템 평가를 위한 테스트셋으로 설정하였다. 설정된 단어들은 정확도 평가의 각 회마다 무작위로 100개씩 추출되었으며, 테스트는 총 10회 진행하였다. 실험을 진행한 결과 본 연구에서 제안한 시스템은 [그림 5]와 같이 모든 실험에 대해 96% 이상의 높은 정확도를 보여주었다.

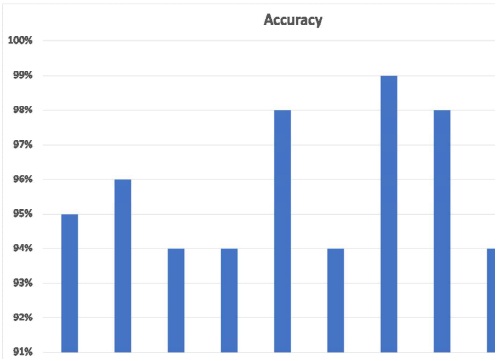

그림 5. 테스트 결과
Note: x-axis: 테스트 회차; y-axis: 정확도(%)

다음으로 본 연구에서 제안하는 LD 알고리즘의 용어 인식 경향은 크게 세 가지로 나누어진다. 첫 번째, LD를 통해 산출된 용어 사이의 거릿값이 최소인 법률 용어로 교정이 이루어지는 경우이다. 예를들어, [표 4], 분류 1의 '단기소멀시효'라는 용어는 '단기소멸시효'와 최소 거리 (1)를 가지기 때문에 교정이 잘 이루어 진다. 두 번째, 입력된 용어와 데이터 베이스 안에 존재하는 용어들 사이에 유사성이 존재하는 경우이다. 예를 들어, [표 4], 분류 2의 '가산새'라는 용어는 '가산세'를 잘못 입력한 사례이다. 하지만 본 연구에서 사용한 데이터 베이스 안에 목표 법률 용어인 '가산세(최소거리 1)'뿐만 아니라 일반 용어인 '가신새(최소거리 1)'도 존재하기 때문에 변환된 음절의 순서가 더 앞에 교정된 '가신새'가 입력된 '가산새'의 교정 결과로 산출 되었다. 마지막으로 입력된 오탈자와 일치하는 용어가 데이터 베이스에 존재하는 경우이다. 예를 들어 [표 4], 분류 3의 '자본제'라는 용어는 산업 재산을 나타내는 법률용어인 '자본재'에 대한 오탈자이다. 하지만 데이터 베이스 안에 '자본 주의 제도'를 뜻하는 일반용어 '자본제'가 있어 시스템이 해당 용어를 오탈자로 인식하지 못하고 교정이 이루어지지 못하였다.

표 4. 오탈자 교정 결과 종류

오탈자	교정 결과	목표한 교정 결과	분류
단기소멀시호	단기소멸시호	단기소멸시효 (성공)	1
난민안전증명서	난민인정증명서	난민인정증명서(성공)	1
가격조시	가격조사	가격조사 (성공)	2
가산새	가신새	가산세 (실패)	2
가암류	가압류	가압류 (성공)	2
발생주의	발생주의	발신주의 (실패)	3
자본제	자본제	자본재 (실패)	3

2. 사용자 평가

본 연구에서는 앞서 수행된 시스템의 정확도 평가와 함께 사용자들을 대상으로 실제 법률 용어를 사용할 때 본 시스템의 사용성에 대한 평가를 진행하였다.

2.1 설문 문항 및 신뢰도 분석

[표 5]는 본 연구에서 사용된 평가 항목과 문항에 대해서 나타내고 있다. 사용자 평가를 위해 본 연구는 선행연구를 기반으로 총 4개의 시스템 평가 항목을 선정하였으며 항목별로 2개의 설문 문항을 설계하였다. 또한 실제 설문을 진행하기에 앞서 일부 표본을 대상으로 설문을 진행하고 설문의 내적 일관성을 확인하기 위한 신뢰도 분석을 실시하였다. 크론바하 알파 계수는 0에서 1사이로 값이 도출되며 1에 가까울수록 내적 일관성이 높다고 할 수 있다. 크론바하 알파 계수를 측정한 결과, 본 실험에 사용될 설문지에 대한 크론바하 알파 계수가 0.851으로 높게 측정되어 높은 내적 일관성으로 설문을 진행하기에 문제가 없다는 점을 확인하고 본 실험을 진행하였다.

표 5. 사용자 평가 문항

항목	문항	선행연구	척도
정확성	1. 법률 용어 자동 교정의 정확도가 좋다고 생각하십니까?	변길현 외(2014), 곽호완 외(2000), 안효선 외(2017), 김상현 외(2011), 임창우 외(2013), 오은혜(2014), 이혜민 외(2014), 최유진(2020)	리커트 5점 척도
	2. 시스템의 교정 결과에 오류가 없다고 생각하십니까?		
용이성	3. 법률 용어 교정 결과에 대한 이해가 쉽다고 생각하십니까?		
	4. 법률 용어 교정에 사용자가 본 시스템을 사용한다면 편리할 것이라고 생각하십니까?		
정보성	5. 시스템이 정확한 법률 용어를 제공한다고 생각하십니까?		
	6. 결과가 세부적으로 도출된다고 생각하십니까?		
사용성	7. 본 시스템을 사용할 의도가 있습니까?		
	8. 본 시스템을 지인들에게 추천할 의향이 있습니까?		

2.2 자료수집 및 표본설정

실험의 집단으로 일반 사용자들을 선정하였으며 실험은 각각 정확성, 용이성, 정보성, 사용성에 대한 문정으로 구성된 실험으로 리커트 5점 척도를 사용하였다. 실험을 위해 일반 사용자 36명을 모집하였으며 온라인 설문지를 배포하여 실험을 진행하였다. 본격적인 사용자 평가를 진행하기 전 법률 용어 사용에 대한 전문성을 탐색하기 위한 설문을 진행하였는데, 법률과 관련된 직접적, 간접적인 경험을 묻는 항목에서 전혀 아니다 7명(19%), 아니다 8명(22%), 보통 8명(22%), 그렇다 12명(33%), 매우 그렇다 1명(3%)으로 나타났고, 법문을 읽거나 작성하려는 경우 법률 용어가 어려움을 묻는 항목은 전혀 아니다 0명(0%), 아니다 2명(6%), 보통 5명(14%), 그렇다 15명(42%), 매우 그렇다 14명(39%)로 나타났다.

실험에 참여한 표본의 인구 통계적 특징은 [표 6]와 같다. 실험에 참여한 총원은 36명이었으며, 성별은 남자가 24명(68.5%), 여자가11명(31.4%)로 나타났다. 연령대는 20대가 31명(88.5%)로 제일 많았고 이어서 50대 이상 3명(8.5%) 40대 1명(2.8%)순으로 나타났다. 학력은 대학교 재학 25명(71.4%) 대학교 졸업 9명(25.7%) 그리고 고등학교 졸업 1명(2.8%)순으로 나타났다.

표 6. 사용자 평가 인구 통계

구분		빈도	비율(%)
성별	남자	24	68.5
	여자	11	31.4
연령	10대 이하	0	0
	20대	31	88.5
	30대	0	0
	40대	1	2.8
	50대 이상	3	8.5
학력	중졸	0	0
	고졸	1	2.8
	대학 재학	25	71.4
	대학 졸업	9	25.7
	대학원 재학 혹은 졸업	0	0

2.3 사용자 평가 결과

[그림 6]은 사용자 평가에 대한 결과로 '전혀 아니다'는 진한 주황색, '아니다'는 연한 주황색, '보통'은 회색, '그렇다'는 연한 초록색, '매우 그렇다'는 진한 초록색으로 표현했다. 시스템에 대한 4개의 항목(정확성, 용이성, 정보성, 사용성)에 대해 평가 결과는 다음과 같다. 우선, '그렇다'와 '매우 그렇다'처럼 긍정적인 응답의 비중이 가장 높은 문항은 용이성에 대한 4번 문항(법률 용어 교정에 사용자가 본 시스템을 사용한다면 편리할 것이라고 생각하십니까?)이었다. 이와 반대로, '전혀 아

니다'와 '아니다'처럼 부정적인 응답의 비중이 가장 높은 문항은 정확성에 대한 2번 문항(시스템의 교정 결과에 오류가 없다고 생각하십니까?)이었다. 이를 통해 본 연구의 시스템이 사용자들이 법률 용어를 사용하는 것을 용이하게 하지만 중요한 법률 용어인 만큼 앞서 96%로 측정된 시스템의 정확성에도 불구하고 더 높은 정확성을 보여주어야 한다는 것을 확인할 수 있었다.

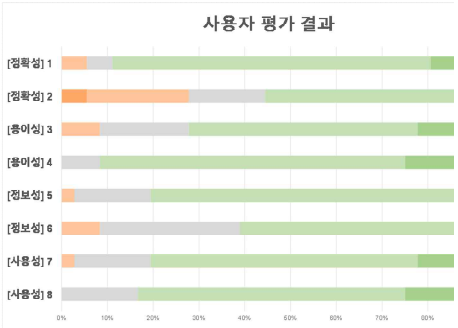

그림 6. 사용자 평가 결과
Note: x-axis: 평과 결과 비율(%); y-axis: 평가 문항 번호

추가로 사용자 평가 이후 시스템에 대한 논의에서는 오류가 0%가 되어야 믿을 수 있겠다는 의견이 있었고, 문맥을 고려하여 교정해주었으면 좋겠다는 의견이 있었다. 그리고 오탈자와 더불어 자동완성기능이 있었으면 좋겠다는 의견 등이 있었다.

V. 논의

본 연구는 LD알고리즘을 활용하여 법률용어의 오탈자를 교정하는 시스템을 제안하였다. 본 연구에서는 제안한 시스템의 사용성을 평가하기 위해 사용자 평가와 10번의 테스트를 진행하였고 그 결과, 제안하는 시스템은 좋은 사용성과 높은 정확도라는 결과를 얻었다. 하지만 이런 긍정적인 결과에도 불구하고 본 연구는 다음과 같은 한계점을 가지고 있다.

우선, 본 연구는 LD 알고리즘을 활용하여 법률용어에서의 오탈자를 교정하는 시스템을 제안했지만, 최근에는 LD 알고리즘처럼 직관적인 알고리즘만 활용하는 것이 아니라 인공지능 신경망과 LD를 병합하여 오탈자를 수정하는 연구들이 제시되고 있다. 예를 들어 폴란드의 Piotr Wojcicki 외(2024)는 CNN(Convolutional Neural Network)모델과 n-gram모델 그리고 LD를 결합한 알고리즘을 제안하였다. 해당 방법론은 오탈자를 교정하기위해 n-gram기반 모델을 사용하여 문자열 사이의 유사도를 계산하고, LD를 활용하여 계산된 거릿값을 종합하여 7개의 단어가 포함된 교정 후보 리스트를 얻는다. 후보 리스트에 있는 7개의 단어 중 오탈자의 위치에 가장 적합한 단어로 교정한다. 본 연구의 경우, LD알고리즘을 한국어에 적용하여 오탈자를 교정하는 과정에서 입련된 용어를 자모로 분리하여 적용하는 등 한국어 분석에 적합한 방법을 제안하였다는데 연구의 의의를 가진다. 하지만 본 연구도 향후 연구를 통해 [34]의 방법론과 같이 인공지능 알고리즘을 병합하여 활용하면 오탈자 교정에 있어 더 높은 정확도가 도출될 것으로 사료된다. 다음으로, [표 4]의 분류 2, 3과 같은 경우에 실패할 확률이 높아지는데, 이를 보완하기 위해서는 선행연구 [9]의 방법론을 생각해 볼 수 있다. [9]는 음성인식에 사용되는 확률적 모델인 히든 마르코프 모델(HMM)에서 인식률을 높이기 위해 교정 결과 후보 어휘 중 보다 적합한 용어를 추출하기 위한 방법론을 제시하였다. 교정 결과 후보 어휘들간의 순서를 정하기 위해서 각 어휘들의 사용 빈도에 따라 가중치를 부여하는 방식이다. 본 연구도 향후 [9]처럼 용어들 간의 순서를 정하기 위해 학습 데이터에서 단어별 사용 빈도를 가중치로 사용하여 적용하여 용어 간의 순서가 없어서 발생하던 분류 2, 3의 문제를 해결하고자 한다. 마지막으로, 일반적인 맞춤법 검사기와 제안한 시스템을 비교하였을 때, 일반 용어에서는 일반적인 맞춤법 검사기보다 낮은 결과를 보여주지만 법률용어는 높은 결과를 보여주었다. 예시로 낙농진흥법의 용어인 '낙농진흥계획'이라는 단어의 오탈자인 '낭농진흥계획'을 입력하였을 때 '사람인', '네이버', '부산대', '잡코리아'의 맞춤법 검사기는 '장롱진흥계획'이라는 잘못된 결과를 보여줬는데 본 연구에서 제안한 시스템은 알맞게 교정해주는 결과를 보였다. 이는 오탈자 교정에 있어 알고리즘의 성능을 향상하기 위해서는 보다 많은 양의 어휘 정보가 필요하다는 것을 의미하며, 향후에는 법률용어뿐만 아니라 다른 분야의 전문 용어를 본 연구 알고리즘에 적용하여 알고리즘의 성능을 고도화할 필요성이 있다.

VI. 결론

본 논문에서는 사용자에게 보다 정확한 법률 용어를 제공하기 위해 음소 분리가 적용된 LD 알고리즘을 이용한 법률 용어 오탈자 교정 시스템을 제안하였다. 본 연구의 결과는 크게 두 가지로 나누어질 수 있다. 첫째, 본 연구에서 제안한 시스템은 법률 용어 오탈자 교정을 하는데에 있어서 높은 정확도를 보여준다. 본 연구는 시스템 평가를 위해서 대법원 법률 종합 정보 사이트의 단어를 사용했고 각각 랜덤한 단어로 실험을 10번 진행했다. 시스템 정확도를 평가한 결과 시스템 정확도는 평균 96%로 나와서 법률 용어를 개선하는데 있어서 높은 결과를 보였다. 둘째, 본 연구에서는 제안하는 알고리즘을 시스템화하여 사용자 평가를 진행한 결과, 용이성과 사용성 항목에 대해서는 사용자들로부터 높은 결과를 얻었다. 해당 결과는 우리의 시스템에 대한 사용성이 높다는 것을 의미한다. 하지만 정확성의 경우 다른 문항들에 비해 다소 부정적인 결과를 얻었는데 향후 시스템의 정확도에 대한 추가 연구를 진행하여 해당 부분에 대한 개선을 진행하고자 한다. 본 연구는 어려운 법률 용어를 일반인들도 쉽게 사용하기 위한 오탈자 교정을 목적으로 진행되었으며, 일반적인 용어의 오탈자 교정과 더불어 법률 용어의 오탈자를 중점적으로 교정하기 위해 많은 양의 법률 용어 데이터를 수집하여 학습시켰다. 진행된 결과 본 연구에서 제안하는 알고리즘의 정확도 및 사용성이 높은 것으로 나타났다. 이는 사용자들이 법률 용어를 사용하는데 있어서 본 연구의 결과가 좋게 활용될 수 있다는 것을 의미한다. 향후 연구에서는 본 연구의 사용자 평가에서 실험한 사용자들의 의견을 기반으로 문맥을 파악해 오탈자를 교정할 수 있도록 개선하고자 하며 본 연구에서 제안하는 시스템이 앞으로 사용자들이 법률 용어를 쉽게 사용하는 데에 있어 더욱 기여하고자 한다.

참 고 문 헌

[1] http://www.kyeongin.com/main/view.php?key=20211123010004462, 2024.2.22.

[2] https://www.lawtimes.co.kr/news/55888, 2024.2.22.

[3] https://scourt.go.kr/portal/news/NewsViewAction.work?seqnum=62&gubun=6&searchOption=&searchWord=, 2024.2.22.

[4] http://www.lec.co.kr/news/articleView.html?idxno=27635, 2024.2.22.

[5] 김지현, 이종서, 이명진, 김우주, 홍준석, “법령정보 검색을 위한 생활 용어와 법률 용어 간의 대응관계 탐색 방법론,” 지능전보연구, 제18권, 제3호, pp.137-152, 2012.

[6] 한애라, “법률문서 자동작성 서비스의 규율에 관한 연구,” 민사소송, 제24권, 제3호, pp.391-449, 2020.

[7] Yulianto, Muhamad Maulana, Riza Arifudin, and Alamsyah Alamsyah, “Autocomplete and spell checking levenshtein distance algorithm to getting text suggest error data searching in library,” Scientific Journal of Informatics, Vol.5, No.1, p.75, 2018.

[8] 안광모, 김윤석, 김영훈, 서영훈, “Levenshtein 거리를 이용한 영화평 감성 분류,” 디지털콘텐츠학회논문지, 제14권, 제4호, pp.581-587, 2013.

[9] 이종섭, 오상엽, “개선된 Levenshtein Distance 알고리즘을 사용한 어휘 탐색 시스템,” 디지털융복합연구, 제11권, 제11호, pp.367-372, 2013.

[10] 이명훈, 양형정, 김수형, 이귀상, 김선희, “간판영상에서 한글 인식 성능향상을 위한 가중치 기반 음소 단위 분할 교정,” 한국콘텐츠학회논문지, 제12권, 제2호, pp.105-115, 2012.

[11] 김은희, 정영미, “사용자 태그와 중심성 지수를 이용한 블로그 검색 성능 향상에 관한 연구,” 정보관리학회지, 제27권, 제1호, pp.61-77, 2010.

[12] 임창우, 주정아, “소셜커머스의 가입여부에 영향을 미치는 요인의 탐색,” 대한경영학회지, 제26권, 제3호, pp.635-671, 2013.

[13] 최유진, 이캐시연주, “패션 4차 산업의 액티브 시니어 고객 유치를 위한 모바일 쇼핑 서비스 개선에 관한 연구 -홈앤쇼핑 애플리케이션의 사용성 평가를 중심으로-,” 한국디자인문화학회지, 제26권, 제1호, pp.511-523, 2020.

[14] 오은혜, “소셜커머스의 구매의도에 영향을 미치는 소셜커머스의 특성 및 관계 품질에 관한 연구,” e-비즈니스연구, 제15권, 제1호, pp.255-275, 2014.

[15] 김상현, 박현선, 김근아, “소셜커머스 특징과 개인 특징이 신뢰와 신뢰성과에 미치는 영향에 대한 실증연

구," 경영연구, 제26권, 제3호, pp.95-121, 2011.
[16] 이혜민, 김승인, "음성인식 기반의 모바일 지능형 개인비서 서비스 사용성 비교," 디지털디자인학연구, 제14권, 제1호, pp.231-240, 2014.
[17] 이현주, "≪우리말샘≫ 편찬 경과," 새국어생활, 제26권, 제4호, pp.65-85, 2016.
[18] 곽호완, 곽지은, 김수진, 이정모, "국내 웹 사이트 디자인의 사용성 조사," 인지과학, 제11권, 제1호, pp.33-45, 2000.
[19] 권기원, 노정란, "2차 법률정보 전문 데이터베이스 구축에 관한 1차 연구," 한국문헌정보학회지, 제32권, 제3호, pp.281-296, 1998.
[20] 변길현, 이혜진, 강신겸, "미술관 관람객의 서비스품질 인식과 만족도 분석 : 광주시립미술관을 중심으로," 문화경제연구, 제17권, 제2호, pp.137-159, 2014.
[21] 정용기, *대국민 대상 정보시스템의 사용자 편의성 개선을 위한 정보제공 최적 경로 사례 연구,* 경북대학교, 국내석사학위논문, 2023.
[22] 박소연, 이준호, "로그 분석을 통한 이용자의 웹 문서 검색 형태에 관한 연구," 정보관리학회지, 제19권, 제3호, pp.111-122, 2002.
[23] 고유강, "법관업무의 지원을 위한 머신러닝의 발전상황에 대한 소고," LAW & TECHNOLOGY, 제15권, 제5호, pp.3-17, 2019.
[24] 김지영, 한다현, 김종권, "빅데이터 검색 정확도에 미치는 다양한 측정 방법 기반 검색 기법의 효과," 정보과학회논문지, 제44권, 제5호, pp.553-558, 2017.
[25] 안효선, 박민정, "소셜미디어 텍스트마이닝을 통한 패션디자인 사용자 인식 조사," 한국의류학회지, 제41권, 제6호, pp.1060-1070, 2017.
[26] 김정민, 황용석, "알고리즘 기반 자동 추천 검색어의 표현물적 특성과 법적 쟁점 -관련 해외 판결을 중심으로-," 언론과법, 제18권, 제2호, pp.1-32, 2019.
[27] 이영신, 박영자, 송만석, "오류 견고성을 지닌 형태소 분석기와 공기정보를 이용한 자동철자 교정," 한국정보과학회 학술발표논문집, 제25권, 제1(B)호, pp.411-413, 1998.
[28] 박광길, 최윤, "우리말샘(개방형 한국어 지식 대사전)을 활용한 신어 연구," 인문과학연구, 제52권, pp.243-266, 2017.
[29] 윤성희, "웹기반 정보검색을 위한 자연어 키워드 색인에 관한 연구," 컴퓨터산업학회논문지, 제4권, 제12호, pp.1103-1111, 2003.
[30] 강승식, 장병탁, "음절 특성을 이용한 범용 한국어 형태소 분석기 및 맞춤법 검사기," 정보과학회논문지(B), 제23권, 제5호, pp.530-539, 1996.
[31] 전정현, 김병필, "인공지능과 법률 서비스 : 현황과 과제," 저스티스, 통권, 제170-1호, pp.218-258, 2019.
[32] 윤성희, "정보 검색 시스템의 성능 향상을 위한 구문 분석과 검색어 확장," 한국산학기술학회 논문지, 제5권, 제4호, pp.303-308, 2004.
[33] 노영훈, 장태우, 원종운, "양방향 RNN과 학술용어사전을 이용한 영문학술문서 교정 방법론," 한국전자거래학회지, 제27권, 제2호, pp.175-192, 2022.
[34] Piotr Wojcicki, and Tomasz Zientarski, "Polish Word Recognition Based on n-Gram Methods," IEEE Access, Vol.12, pp.49817-49825, 2024.
[35] Putra, Made Edwin Wira, and Iping Supriana Suwardi, "Structural off-line handwriting character recognition using approximate subgraph matching and levenshtein distance," Procedia Computer Science, Vol.59, pp.340-349, 2015.
[36] https://www.klri.re.kr/kor/business/bizLawDicKeyword.do, 2024.2.22.
[37] https://glaw.scourt.go.kr/wsjo/intesrch/sjo022.do
[38] https://www.moleg.go.kr/mpbleg/mpblegInfo.mo?mid=a10402020000&searchCondition=CTS&searchKeyword=%EC%8B%A0%EA%B3%A0&pageIndex=38&mpb_leg_pst_seq=131058, 2024.2.22.

저 자 소 개

공 성 호(Sung-Ho Gong) 준회원

■ 2018년 3월 ~ 현재 : 아주대학교 사학과 학사과정

〈관심분야〉 : 디지털인문학, 자연어처리, 데이터분석

심 진 우(Jin-Woo Shim) 준회원

■ 2022년 3월 ~ 현재 : 아주대학교 문화콘텐츠학과 학사과정

〈관심분야〉 : 문화콘텐츠학, 디지털인문학, 데이터시각화

현 민 호(Min-Ho Hyeon) 준회원

■ 2020년 3월 ~ 현재 : 아주대학교 사학과 학사과정

〈관심분야〉 : 역사학, 디지털인문학, 데이터시각화, 통계학

문 성 민(Seongmin Mun) 정회원

■ 2021년 6월 : Université Paris Nanterre Sciences du Langage (전산언어학 박사)

■ 2021년 8월 ~ 2022년 10월 : 조선대학교 영어영문학과 박사후연구원

■ 2022년 11월 ~ 현재 : 아주대학교 인문과학연구소 연구교수

〈관심분야〉 : 전산언어학, 디지털인문학, 자연어처리, 통계분석, 기계학습, 신경망분석, 인공지능, 데이터시각화