https://doi.org/10.5392/JKCA.2020.20.10.025

질의응답 커뮤니티에서 문서 간 이독성 비교

Comparison of Readability between Documents in the Community Question-Answering

문길성
국민연금공단 정보전략실

Gil-Seong Mun(gsmun@nps.or.kr)

요약

커뮤니티 질의응답 서비스는 다양한 목적으로 활용되고 있다. 질의응답 문서에서 정보의 품질은 질문의 명확성과 답변 내용의 적절성으로 결정되며 문서의 읽기 쉬운 정도를 나타내는 이독성(readability)은 문서가 가지고 있는 정보의 품질을 평가하기 위한 주요 요소이다. 본 연구의 목적은 국내의 CQA 사이트에서 제공되는 문서의 품질을 측정하는 것이다. 이를 위하여 네이버 지식iN의 '국민 신문고' 커뮤니티의 문서에서 사용된 어휘 수준별 사용 빈도를 비교하고, 작성 기관별 문서의 이독성 지수를 측정하였다. 이독성 지수의 측정은 어휘 수준과 문장 길이를 바탕으로 계산되는 Dale-Chall 공식을 사용하였다. 분석 결과, 답변에서 사용된 어휘는 질문에서 사용된 어휘보다 더 어려운 수준이고 문장 길이도 더 길어서 이독성이 더 낮은 것으로 나타났다. 또한, 질의응답간 이독성의 차이는 작성 기관별로도 차이가 있음을 파악할 수 있었다. 본 연구의 결과는 상담 업무에 반영할 수 있는 유용한 정보를 도출하여 온라인상의 민원상담 개선에 활용할 수 있으며, 이독성 지수에 기반하여 문서 수준의 정량적 분석을 시도함으로써 텍스트 마이닝의 주제를 확장할 수 있을 것으로 기대한다.

■ 중심어 : | 커뮤니티 질의응답 사이트 | 이독성 | 이독성 공식 | 텍스트 마이닝 |

Abstract

Community question and answering service is one of the main sources of information and knowledge in the Web. The quality of information in question and answer documents is determined by the clarity of the question and the relevance of the answers, and the readability of a document is a key factor for evaluating the quality. This study is to measure the quality of documents used in community question and answering service. For this purpose, we compare the frequency of occurrence by vocabulary level used in community documents and measure the readability index of documents by institution of author. To measure the readability index, we used the Dale-Chall formula which is calculated by vocabulary level and sentence length. The results show that the vocabulary used in the answers is more difficult than in the questions and the sentence length is longer. The gap in readability between questions and answers is also found by writing institution. The results of this study can be used as basic data for improving online counseling services.

■ keyword : | Community Question-Answering | Readability | Readability Formula | Text Mining |

접수일자 : 2020년 07월 31일
수정일자 : 2020년 09월 15일
심사완료일 : 2020년 09월 17일
교신저자 : 문길성, e-mail : gsmun@nps.or.kr

I. 서 론

웹(Web) 서비스로 대표되는 인터넷 기술은 정보 검색 및 공유 방식을 바꾸어 놓고 있다. 지난 수십 년간 새로운 정보 채널 및 자원(resource)이 급격히 증가함에 따라 사용자들은 정보 생산과 소비의 주체로서 다양한 유형의 온라인 정보원(information source)을 사용하고 있다. 이러한 정보원에는 위키, 포럼, 블로그, 커뮤니티 질의응답(community question answering; CQA) 서비스 등이 존재한다[1]. 특히, CQA 서비스는 2002년 네이버(Naver) 지식iN을 시작으로 Yahoo! Answers, Stack Exchange, Quora 등이 대표적이라 할 수 있으며, 다양한 사용자들이 요구하는 정보를 제공할 수 있는 핵심 서비스로 부상하였다. 이는 자연어 형태의 질의가 가능하고 사용자 개개인에 적합된 맞춤형 답변을 얻을 수 있으며, 기존의 키워드 기반 질의 방식에 비하여 더 정확한 형태로 정보를 얻는데 도움이 되기 때문이다[2].

최근에는 정부부처, 정부기관, 지자체나 공공기관에서 민원상담업무 및 소통 창구로도 CQA를 활용하는 등 그 범위가 확대되고 있다. 이들 기관은 '네이버 지식iN'을 주로 이용하고 있으며 네이버 지식iN은 답변 품질을 높이기 위하여 전문적이고 신뢰성 있는 답변을 공식적으로 작성할 수 있도록 2017년부터 지식파트너[3] 제도를 운영하고 있다. CQA 서비스에 대한 사회적 관심이 커짐에 따라 질문의 명확성이나 답변의 유용성과 같은 콘텐츠의 품질을 평가하고 응답행위 및 답변 채택 여부를 예측하는 연구가 활발하게 수행되고 있다[4-7]. 이들 연구는 공통적으로 얼마나 쉽게 읽을 수 있는가를 나타내는 이독성(readability) 지수, 특히 답변의 이독성 지수를 주요 변수로 고려하고 있으나 질의응답 간 이독성의 차이는 고려하지 않고 있다. 또한, Stack Exchange와 같이 영어로 작성된 문서의 품질 측정에 대한 연구는 많이 이루어지고 있으나 한글로 작성된 문서에서의 품질을 측정한 연구는 미흡한 실정이다.

본 연구의 목적은 국내의 CQA 사이트에서 제공되는 문서의 품질을 측정하는 것이다. 이를 위하여 한글 문서에 적합한 이독성 공식들(readability formula)을 검토하였다. 네이버 지식iN의 '국민 신문고' 커뮤니티의 문서를 분석 대상으로 하여 질의응답에서 사용된 어휘 수준별 사용 빈도를 비교하고, 작성 기관별 문서의 이독성 지수를 측정하였다. 어휘의 수준별 출현 빈도의 시각화를 통하여 질의응답 간 어휘 사용 양상을 비교 분석하였으며, 질의응답 간 이독성 비교를 위해 답변 내용을 작성 기관별로 비교하여 텍스트의 수준 차이를 살펴보았다. 이독성 지수의 계산은 어휘 난이도와 같은 어휘 요인과 문장 복잡도와 같은 통사적 요인이 요구된다. 전처리 과정을 통하여 질의응답 문서들에 대한 형태소 분석 후 문장 복잡도를 계산하고 국립국어원에서 제공하는 한국어기초사전[8]에서 제공되는 어휘 수준을 이독성 공식에 대입하여 지수를 측정하였다.

본 연구의 결과는 상담업무에 반영할 수 있는 유용한 정보를 도출하여 온라인상의 민원상담 개선에 활용할 수 있으며, 토픽모형[9][10]이나 감성분석[11]과 같은 텍스트 마이닝의 대표적 주제 외에 이독성 지수에 기반하여 문서 수준의 정량적 분석을 시도함으로써 텍스트 마이닝의 주제를 확장할 수 있을 것으로 기대한다.

II. 텍스트 난이도와 이독성 공식

1. 개념

텍스트의 난이도(difficulty)나 글의 수준에 대한 평가는 언어학, 교육학, 심리학 등의 다양한 분야에서 오랫동안 연구된 주제이다[12]. 글의 수준을 평가하는 방법은 크게 양적 평가와 질적 평가로 나눌 수 있다. 양적 평가는 어휘 난이도나 문장의 길이와 같은 표면적 특성에 기초하여 글의 수준을 객관적으로 평가하는 방법이다. 텍스트의 양적 측면만을 이용하여 읽기 쉬운 정도를 이독성으로 정의하고 주로 이독성 공식을 사용하여 글의 수준을 평가한다. 이에 반해, 질적 평가는 전문가의 경험이나 직관에 기초하여 글의 수준을 주관적으로 평가하는 방법이다. 주로 글의 구조, 독자의 선행지식이나 흥미 등을 평가 기준으로 삼아 글의 수준을 판단한다[13][14]. 텍스트의 양적, 질적 측면 외에 글자꼴이나 줄 간격 등과 같은 형태적 측면을 강조한 가독성(legibility) 평가도 있으나 본 연구에서 언급하는 이독성과 구분할 필요가 있다. 본 연구는 CQA에서 질의응

답간 텍스트 난이도를 비교 분석하는 데 초점을 두기 때문에 글의 수준을 객관적으로 수치화할 수 있는 양적 개념만을 사용한다.

2. 한국어 텍스트에 적합한 이독성 공식

글의 수준을 정량적인 수치로 나타내고 기계적 계산을 위해서는 측도로서 타당하고 구현이 용이한 이독성 공식이 필요하다. 이독성에 대한 연구는 1920년대 시작하여 현재까지 꾸준히 진행되고 있으며, Flesch, Gunning FOG, SMOG, Flesch-Kincaid, Dale-Chall, Coleman-Liau 지수 등이 대표적이다[15][16].

이독성 공식은 공통적으로 의미적 변수(semantic variable)와 통사적 변수(syntactic variable)로 구성되는데, 전자는 어휘 차원의 변수로 어휘의 길이나 난이도로 측정하며 후자는 문장 차원의 변수로 대부분 문장 길이로 측정한다. 한국어 텍스트의 이독성 공식은 어휘 차원의 변수로 어휘의 난이도만을 고려하는데, 어휘의 길이를 사용하지 않는 이유는 언어 특성상 한자의 영향을 받아 길이가 짧아도 어려운 어휘일 수 있어 이독성을 왜곡시킬 수 있다고 판단하기 때문이다[13][14]. 문장의 길이와 어려운 어휘를 바탕으로 텍스트의 난이도를 측정하는 대표적인 공식으로 Dale-Chall 공식이 있다. DuBay[15]는 영어 이독성 공식 중 Dale-Chall 공식을 가장 타당한 공식이라 인정했으며, 국내에서도 이독성 측정을 위해 가장 많이 사용되고 있다[13][14][17]. 문서 d에 대한 Dale-Chall의 이독성 공식 R(d)는 식 (1)과 같다. 식 (1)에서 words는 주어진 문서에서 사용된 단어 수, difficult words는 어려운 단어 수, sentences는 문서에서 사용된 문장의 수이다. 식의 우변에서 앞부분은 어려운 단어의 비율을 의미하며 뒷부분은 문장 길이를 정의한 것으로 문장당 사용된 단어 수이다.

$$R(d) = 0.1579\left(\frac{difficult\ words}{words} \times 100\right) + 0.0496\left(\frac{words}{sentences}\right) + 3.6365 \quad (1)$$

고승연[14]은 영어권에서 개발된 공식들이 문법적 요소는 고려하지 않는 등 한국어 텍스트의 이독성을 평가하는데 적합하지 않다고 지적하였다. 박정진[18]은 한국어에 특화된 이독성 공식에 관한 연구가 미비하며 개발된 공식들도 초등 텍스트 수준에 국한되어 있고 실질적으로 사용된 사례를 찾기 어렵다고 언급하였다. 따라서 한국어에 적합한 이독성 공식이 앞으로 더 실제적으로 연구되고 개발될 필요가 있으며, 현재는 영어권에서 개발된 공식을 사용하는 것이 일반적이다. 본 연구에서는 CQA에서 질의응답 문서의 이독성 비교를 위하여 어휘 요인과 문장 요인을 고르게 반영하고 있으면서 타당도가 높은 것으로 알려진 Dale-Chall 공식을 활용한다.

3. Dale-Chall 이독성 공식의 활용 방안

한국어 텍스트에 Dale-Chall 공식을 활용하기 위해서는 단어의 수와 어려운 단어를 파악하기 위한 기준이 필요하다. 또한, 주어진 문서가 몇 개의 문장으로 구성되어 있는지 파악할 수 있어야 하며, 이를 위해서는 문장을 구분할 수 있어야 한다. 웹에서 작성된 문서들은 비문이거나 문법적 오류가 많아 단순히 마침표나 물음표와 같은 문장부호 기준으로 문장을 구분할 수 없다는 어려움이 존재한다. 문장 구분을 위하여 필수적으로 전처리 과정이 요구되며 자세한 방안은 3.3절에서 논의하기로 한다.

문장에서 단어의 수를 세기 위해서는 단어들을 개별적으로 구분할 수 있어야 한다. 영어에서는 띄어쓰기로 단어를 구분할 수 있지만, 한국어는 영어와 다르게 실질적 의미를 갖는 단어 또는 어간에 문법적 기능을 가진 요소가 결합되어 있으므로 단순히 띄어쓰기 단위로 단어를 구분하는 것은 적절하지 못하다. 따라서 실질적 의미를 갖는 부분을 단어로 구분하고 이 부분에 해당하는 단어의 품사를 정할 필요가 있다. 본 연구에서는 단어 구분과 개수 파악을 위하여 어미, 조사 등과 같은 관계사류와 실질적인 의미가 없는 기능동사 '이다', 그리고 수사를 제외한 명사, 대명사, 동사, 형용사, 관형사, 부사, 감탄사, 의존 명사, 보조 동사, 보조 형용사와 같은 품사의 단어를 사용한다.

어려운 단어와 쉬운 단어를 구분하기 위하여 영어에서는 쉬운 어휘를 모아 놓은 Dale-Chall 리스트를 기

준으로 목록에 없는 단어를 어려운 단어로 분류한다. 한국어 텍스트의 이독성 공식 개발 연구에서도 이러한 영향을 받아 어휘 목록을 사용한다. 고승연은 한국어 어휘 목록 선정과 관련한 연구들을 체계적으로 정리하면서 한송화[19]가 개발된 어휘 목록의 사용을 추천하였다. 한송화가 개발한 한국어 교육 어휘는 국립국어원의 한국어기초사전에 구현되어 있으며, 초급 1,835개, 중급 3,855개, 고급 4,945개로 총 10,635개를 세 수준으로 분류되어 있다. 초급은 일상적이고 친숙한 상황에서 사용 가능한 기초어휘이며 중급은 일상적이고 사회적 상황에서 확장된 기초어휘이다. 고급은 사회적, 전문적, 친숙하지 않은 어휘들로 구성되어 있다.

Dale-Chall 이독성 공식에 의해 산출된 지수는 [표 1]과 같이 학년 수준으로 변환하여 활용할 수 있다. Dale-Chall 지수가 10 이상이면 대학 졸업자 이상으로 판정하고, 한 학년에 해당하는 지수는 0.5이며 이독성 지수가 높을수록 더 어려운 텍스트인 것으로 해석할 수 있다.

표 1. Dale-Chall 이독성 지수와 학년 수준

Dale-Chall 지수	미국기준 학년 수준	한국기준 학년 수준
4.9 이하	4학년 이하	초등 4학년 이하
5.0~5.9	5~6학년	초등 5~6학년
6.0~6.9	7~8학년	중학교 1~2학년
7.0~7.9	9~10학년	중3~고등학교 1학년
8.0~8.9	11~12학년	고등학교 2~3학년
9.0~9.9	13~15학년 (대학생)	대학생
10.0 이상	16학년 이상	대학 졸업자 이상

III. 자료 설명 및 분석 방법

1. 자료 소개

국내의 CQA 사이트에서 제공되는 텍스트의 질의응답간 이독성 비교를 위하여 온라인 민원상담 자료를 활용하였다. 대표적 민원처리 기관 중 하나인 국민권익위원회는 국민신문고를 운영하고 있으며, 네이버 지식iN에서도 상담한 내역을 제공하고 있다. 본 연구에서 사용된 자료는 국민신문고의 민원상담 자료이며 웹 크롤링(web crawling)을 통하여 수집하였다. 2005년부터 2019년까지 15년간 축적된 89,365개의 민원상담 자료로 질의응답 내용, 작성 시간, 조회 수 등으로 구성되어 있으며 작성 기관은 정부 부처, 교육청, 지자체 등이다.

2. 분석 수행 절차

수집된 자료는 꼬꼬마 형태소 분석기[20]와 R 프로그램을 사용하여 분석하였다. 텍스트 마이닝 기법을 이용하여 보다 더 정확한 분석과 의미 있는 결론을 도출하기 위해서는 텍스트에 포함된 특수문자나 문장부호를 제거하고 띄어쓰기 수정 등의 정제 과정이 필요하다. 문서에서 문장과 단어 분리(구분) 및 추출을 위하여 형태소 분석도 필수적이다. '꼬꼬마 형태소 분석기'는 다른 형태소 분석기에 비해 풍부한 분석 태그를 제시함으로써 세부적인 성분 분석이 가능하며, 품사 태깅의 정확도가 상대적으로 높다는 장점이 있다[21].

형태소 분석 후 질문과 답변에 나타난 단어 수, 문장수, 어려운 단어의 수를 세고 이러한 정보를 바탕으로 식 (1)을 이용하여 이독성 지수를 계산하였다. 또한, 질의응답간 어휘 사용 양상을 관찰하기 위하여 어휘 수준별로 출현 빈도를 구하고 워드클라우드로 시각화하여 어휘들의 분포를 비교하였으며, 작성 기관별로 질의응답 간 이독성을 비교 분석하였다.

3. 전처리 과정

비정형 텍스트 자료를 정형 자료로 변환하기 전에 의미 없고 불필요한 문자와 텍스트에 반복적으로 나타난 상용어구 등을 파악하여 이를 분석에서 제외하는 정제 작업이 필요하다. 웹에서 수집된 자료의 특성상 띄어쓰기 오류를 비롯하여 HTML 태그나 특수문자 등 정확한 분석을 방해하는 요소가 다수 포함되어 있기 마련인데, Java와 같은 프로그래밍 언어에서 일반적으로 제공하는 정규식(regular expression)과 문자열 대체 기능을 이용하여 정제 작업을 진행하였다. 작성 기관별 이독성 비교를 위하여 답변 문서에 포함된 기관 정보를 추출한 후 원자료에서 제거하였다. 텍스트 마이닝 절차상 많은 경우에 불용어(stopword)를 선정하여 제외하나 본 연구에서는 중요 키워드를 찾는 것이 아닌 텍스트의 난이도 측정에 목적을 두고 있기 때문에 불용어 처리는 별도로 하지 않았다.

정제된 텍스트 자료는 형태소 분석을 통하여 품사 태깅(pos tagging)을 하였으며 2.3절에서 언급한 품사의 단어만 선정하였다. 선정된 단어는 품사가 부착된 형태(예를 들면 '지원/NNG', '또한/MAG')로 쉬운 어휘 목록에서 조회하여 분석에 이용하였다. 맞춤법상 오류가 없다면 문장의 끝을 지시하는 문장부호를 이용하여 문장을 구분하고 그 수를 셀 수 있으나 웹 문서 특성상 문법적 오류가 많고 문장부호가 생략된 경우가 많아 본 연구에서는 '평서형 종결 어미'나 '의문형 종결 어미'와 같은 종결 어미를 이용하여 문장을 구분하였다.

마지막으로 쉬운 어휘 목록을 사전 형태로 구축하기 위하여 국립국어원에서 제공하는 한국어기초사전을 이용하였다. 사전 파일을 어휘 수준이 초급, 중급, 고급인 단어만 '단어/품사' 형태를 키(key)로 하고 어휘 수준을 값(value)으로 하는 키-값 쌍의 구조로 변경하여 형태소 분석으로부터 추출된 단어의 어휘 수준을 조회할 수 있도록 하였다. 어휘 목록 작성 과정에서 두 가지 수정 작업이 요구된다. 형태소 분석기에서는 '단어/품사' 형태가 아닌 '단어/어근+단어/접사' 형태가 빈번하게 존재하나 한국어기초사전에는 이 형태를 동사나 형용사 중 하나로 취급하고 있어 형태소 분석기의 출력 형태에 맞게 통합할 필요가 있다. 또한, 접두사와 접미사를 구분하는 형태소 분석기와 달리 한국어기초사전에서는 접사로만 제공하므로 형태소 분석 결과에서 접두사와 접미사를 접사로 통합할 필요가 있다.

IV. 분석 결과

1. 어휘 수준별 분포

[표 2]는 질의응답에서 사용된 어휘 수준들을 비교한 결과이다. 질문과 답변이라는 텍스트 유형과 사용되는 어휘 수준의 연관성을 파악하기 위하여 독립성 검증을 실시한 결과 명사를 제외한 나머지 품사들은 연관성이 없는 것으로 분석되었다. 질의응답 간 명사의 사용 양상을 살펴보면 답변에서 사용된 어휘가 질문에서 사용된 어휘에 비하여 좀 더 어려운 어휘가 사용된 것으로 나타났다. 전체 어휘에서 명사가 차지하는 비율은 대략 90%로 아주 높고, 질문에 비하여 답변에서 어려운 어휘('고급', '등급 외')가 사용되는 비율이 더 높으므로 답변 텍스트의 이독성이 더 낮을 것으로 추측할 수 있다.

표 2. 질의응답에 사용된 어휘 수준 비교

품사	어휘 수준	질의	응답	χ^2
명사	초급	1,023 (4.0)	1,039 (2.8)	399.37***
	중급	2,186 (8.6)	2,252 (6.0)	
	고급	2,578 (10.2)	2,852 (7.7)	
	등급 외	19,557 (77.2)	31,132 (83.5)	
동사	초급	253 (14.7)	250 (14.9)	3.68
	중급	478 (27.7)	488 (29.1)	
	고급	278 (16.1)	295 (17.6)	
	등급 외	716 (41.5)	645 (38.4)	
형용사	초급	94 (25.9)	89 (28.1)	0.63
	중급	61 (16.8)	54 (17.0)	
	고급	55 (15.2)	43 (13.6)	
	등급 외	153 (42.1)	131 (41.3)	
부사	초급	126 (12.3)	123 (14.0)	2.43
	중급	237 (23.1)	205 (23.3)	
	고급	196 (19.1)	148 (16.8)	
	등급 외	469 (45.6)	405 (46.0)	

p<.05: *, p<.01: **, p<.001: ***

[그림 1]은 질문과 답변에서 사용된 어휘 중 품사가 명사인 단어들의 전체적인 분포를 워드클라우드로 나타낸 것이며, [그림 2]는 어휘 수준별로 질문과 답변에 사용된 단어의 분포를 비교한 것이다. [그림 1]을 살펴보면 질문과 답변에서 '초급'과 '중급' 수준의 단어(쉬운 어휘)가 '고급'과 '등급 외' 수준의 단어(어려운 어휘)에 비하여 출현 빈도가 확연히 높다는 것을 알 수 있다. 특히 질문에서 사용된 쉬운 어휘들의 출현 빈도가 답변에서 사용된 어휘에 비하여 높게 나타나고 있다. 한편 [그림 1]의 우측 워드클라우드에서 어휘 수준이 등급 외인 경우 '법령'이라는 단어가 두드러지게 나타나고 있는데 이는 답변을 작성하는 기관에서 민원상담 특성상 관련 근거로 법령을 많이 인용하기 때문이다.

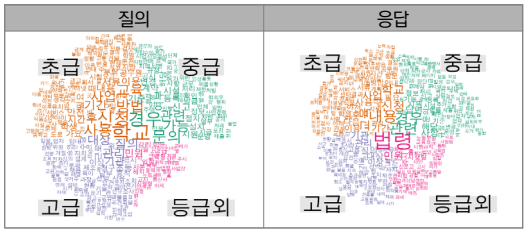

그림 1. 질의응답에 대한 어휘 수준별 워드클라우드

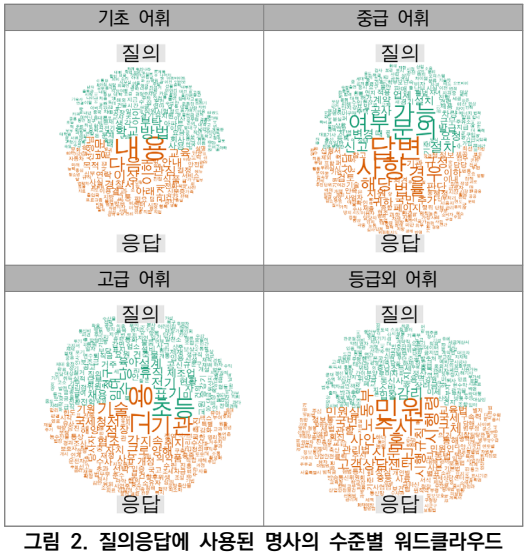

그림 2. 질의응답에 사용된 명사의 수준별 워드클라우드

[그림 2]에서 어휘 수준별로 질의응답간 사용된 단어들을 살펴보면 먼저, 초급 수준에서는 질문의 경우 '방법', '부탁', '학교', '사용'과 같은 단어의 출현 빈도가 높은 반면 답변에서는 '내용', '다음', '이해', '관심', '방문'과 같은 단어의 출현 빈도가 높게 나타났다. 중급 수준에서는 '가능', '여부', '문의', '신고', '절차' 등의 단어가 질문에서 높게 출현 되었고 '답변', '사항', '법률', '해당', '규정' 등의 단어가 답변에서 높게 출현된 것으로 분석되었다. 고급 수준 어휘로 질문에서는 '초등', '표기', '임대', '육아', '휴직', '청구' 등의 단어가 출현 빈도가 높았고, 답변에서는 '고용', '근로', '기관', '기술', '양해', '취지' 등이 높게 나타났다. 마지막으로 '등급 외'의 어휘 수준에서는 답변의 '법령'을 제외하면 두드러지게 나타나는 단어가 없어 '법령'을 제외한 후 워드크라우드를 다시 작성하였다. 그 결과 답변에서 사용된 단어 수가 질문에서 사용된 단어 수에 비하여 훨씬 많고 출현 빈도가 높은 단어도 많은 것으로 파악되었다. 등급 외의 어휘 사용 양상을 보면 질문에서는 '감리', '입찰', '소방', '통신', '위약금' 등의 출현 빈도가 높았고, 답변에서는 '민원', '주시', '시행령', '시행규칙', '홈' 등의 출현 빈도가 높게 나타났다.

어휘 수준별로 질문과 답변에서 사용된 단어들을 살펴본 결과 공통적으로 출현 빈도가 높은 단어는 나타나지 않아 질문과 답변에서 사용되는 단어가 상이하다는 것을 알 수 있다. 또한, 질문의 주요 핵심 단어가 방법, 절차, 이유 등으로 구성된다면 답변은 질문에 대한 설명, 이해, 근거가 주를 이루고 있다는 사실을 파악할 수 있다.

2. 작성 기관별 이독성 비교

[표 3]은 질문과 답변에 대한 문장 요인과 어휘 요인의 비교 결과이다. 작성 기관의 수가 너무 많아 상위 부서나 기관으로 다시 범주를 재조정하였으며 자료 건수가 100개 이상인 기관만 표에 제시하였다. 기관별 자료 건수를 볼 때 교육청, 행정안전부, 광역지자체, 해양수산부, 기획재정부의 민원상담 건수가 다른 기관에 비하여 상대적으로 많으며 전체 자료의 66.4%를 차지하고 있다.

전반적으로 답변의 문장 길이가 질문의 문장 길이보다 13 단어 정도 길고 어려운 단어의 비율은 1%p 정도 높은 것으로 분석되었다. 이러한 결과를 식 (1)에 적용할 때 어휘 요인은 0.1579, 문장 요인은 0.6448의 차이를 만들어 문장 길이가 이독성 지수에 4배 정도 더 영향을 주며 한 학년 이상의 차이를 만드는 것으로 파악되었다. 이는 이독성이 높은 답변을 위해서는 이해하기 쉬운 어휘를 사용하는 것도 중요하지만 문장을 짧게 작성하여 문장 길이를 줄일 필요가 있음을 의미한다.

기관별로 문장 요인을 비교하면 중소벤처기업부, 금융위원회, 농림식품부는 답변의 문장 길이가 질문보다 압도적으로 길게 나타난 반면 산업통상자원부, 식품의약품안전처는 질문의 문장 길이가 더 짧은 것으로 나타났다. 중소벤처기업부가 답변에서의 문장 길이가 비정상적으로 길게 나타난 이유는 문장의 수가 평균 2개로 다른 기관에 비하여 적기 때문이며 원자료를 검토했을 때 동사나 형용사와 같은 용언의 사용이 거의 없는 업

표 3. 질의응답 문서에서 평균 문장 길이와 어려운 어휘의 비율

작성 기관	건수	질의					응답				
		문장수	단어수	어려운 단어수	문장길이	어려운 단어비율	문장수	단어수	어려운 단어수	문장길이	어려운 단어비율
전체	89,236	2.1	31.9	11.2	15.9	35.0	5.7	124.8	44.7	29.0	36.0
교육청	23,513	1.8	24.7	7.7	14.1	31.1	4.1	92.7	31.0	27.4	33.7
행정안전부	13,641	2.7	48.2	17.0	18.5	33.6	7.3	140.4	48.1	22.1	34.2
광역지자체	7,655	1.8	22.6	7.7	12.3	34.7	4.4	114.5	40.7	39.3	35.9
해양수산부	7,366	1.5	17.7	7.1	11.7	41.6	4.4	101.6	42.7	27.5	42.6
기획재정부	7,117	2.1	28.9	10.9	14.6	38.2	5.6	113.5	43.8	24.4	39.0
고용노동부	4,178	1.4	21.3	7.7	16.6	37.7	8.4	165.2	61.4	23.2	37.7
농림축산식품부	3,511	2.6	37.2	14.2	16.0	38.0	5.7	166.4	66.2	59.9	40.0
산업통상자원부	3,403	2.7	71.3	26.8	35.2	37.1	6.1	158.4	58.4	28.9	36.8
과학기술정보통신부	3,161	2.8	41.1	14.4	16.5	35.8	8.3	146.9	44.9	18.0	30.5
국방부	2,995	2.2	29.8	10.9	14.7	37.8	5.5	126.3	48.4	28.0	38.6
식품의약품안전처	1,899	1.5	31.2	12.3	21.5	39.1	8.4	147.0	54.5	17.4	37.2
문화체육관광부	1,828	1.9	20.9	7.4	11.2	38.1	7.6	105.7	37.1	14.1	35.1
환경부	1,801	1.2	16.9	6.3	14.1	37.5	4.0	92.2	35.3	31.6	37.8
방송통신위원회	1,494	2.4	39.9	12.2	19.9	30.0	8.4	257.8	85.8	39.5	32.7
법무부	1,328	5.0	75.9	26.6	16.8	33.8	11.6	237.1	87.2	19.6	34.8
인사혁신처	1,027	2.5	45.4	15.3	20.5	34.1	4.4	108.0	37.3	24.7	34.4
중소벤처기업부	797	1.1	9.1	3.2	8.1	36.6	2.0	244.1	79.5	133.5	33.1
공정거래위원회	774	2.6	46.3	16.9	20.7	36.4	5.2	148.9	53.9	36.1	37.4
금융위원회	330	1.1	17.6	5.0	17.0	28.7	1.6	79.8	28.1	62.1	35.0
국가보훈처	301	2.2	31.2	11.4	14.9	39.0	8.3	181.9	67.7	21.9	37.2
보건복지부	238	1.3	12.0	4.3	10.2	36.4	4.6	95.6	34.9	20.7	36.4
교육부	185	1.7	30.3	8.6	22.0	26.9	3.2	76.5	22.9	26.8	31.5
국토교통부	162	1.4	18.2	5.0	14.6	26.8	4.4	91.5	29.8	24.5	33.0
법제처	149	1.5	25.0	9.7	15.8	39.6	5.0	116.8	43.8	28.1	37.5
외교부	142	3.9	50.3	14.4	13.8	28.4	9.0	147.2	45.9	18.5	30.8
한국산업인력공단	136	1.4	19.9	4.3	15.5	21.1	2.8	53.0	16.4	20.0	31.6
통일부	105	2.1	26.0	8.6	16.4	34.6	8.4	165.0	53.0	21.4	31.3

무 지침이나 안내서 등을 그대로 복사하여 붙인 자료가 많기 때문인 것으로 조사되었다. 기관별로 어휘 요인을 비교할 때 한국산업인력공단, 금융위원회, 국토교통부, 교육부 등은 답변에서의 어려운 어휘 비율이 질문보다 높게 나타났고, 과학기술정보통신부, 중소벤처기업부, 통일부는 질문에서의 어려운 어휘 사용 비율이 답변보다 상대적으로 높은 것으로 나타났다.

[표 4]는 답변 작성 기관별로 질문과 답변의 Dale-Chall 이독성 지수를 비교한 결과이다. 질문에 대한 이독성 지수의 전체 평균은 10.0, 답변은 10.8로 평균적으로 0.8 만큼 답변의 이독성 지수가 높은 것으로 나타났다. 한 학년에 해당하는 지수는 0.5 정도이므로 답변이 질문보다 두 학년 가까이 어려운 수준으로 작성된 것으로 해석되며, 이러한 차이는 [표 3]의 결과를 종합해볼 때 문장의 길이가 주요 원인 것으로 판단할 수 있다. 이독성 지수가 10 이상이면 대학교 졸업 이상의 학력을 갖는 것으로 추정되는데 이는 국민신문고에 민원상담을 하는 계층이 성인이며, 학력 수준이 높다는 특성이 반영된 결과로 해석된다. 질의응답 간 이독성 차(답변 - 질문)가 큰 기관, 즉 질문과 비교해 답변이 읽기 어렵게 작성된 기관은 [표 3]의 결과에서 예상했듯이 중소벤처기업부, 금융위원회, 농림축산식품부 등이고 과학기술정보통신부, 식품의약품안전처 등인 것으로 파악되었다.

[표 4]의 결과에서 질의응답 간 이독성의 차이와 더불어 이독성에 대한 최댓값과 왜도를 주목할 필요가 있다. 답변 작성의 개선을 위해서는 특히 답변의 최댓값과 왜도를 주목할 필요가 있는데, 비정상적으로 이독성 지수가 큰 답변은 더욱 읽기 쉬운 형태로 작성할 필요성이 있는 텍스트이기 때문이다. 농림축산식품부의 최대 이독성은 130.1이며, 원자료를 검토한 결과 '인삼특용작물계열화(융자) 사업'에 대한 질문에 사업 계획(안내서) 자체를 답변으로 제공해 발생한 것으로 파악되었다. 다시 말하면 질문 내용에 맞게 답변을 제공하기보다는 이미 작성된 문서를 수정 없이 재사용하면서 발생한 문제라 할 수 있다.

표 4. 질의응답 문서쌍에서 Dale-Chall 이독성 비교

작성 기관	건수	질의				응답				평균 이독성 차이
		최솟값	최댓값	왜도	평균	최솟값	최댓값	왜도	평균	
전체	89,236	3.7	30.0	0.2	10.0	3.8	130.1	5.5	10.8	0.8
교육청	23,513	3.7	19.7	0.2	9.3	4.9	51.2	2.4	10.3	1.1
행정안전부	13,641	3.7	30.0	0.3	9.9	4.9	36.1	2.0	10.1	0.3
광역지자체	7,655	3.7	19.8	0.4	9.7	3.8	46.3	2.9	11.3	1.5
해양수산부	7,366	3.7	20.0	0.1	10.8	5.2	36.3	2.5	11.7	1.0
기획재정부	7,117	3.7	19.7	0.2	10.4	5.6	30.0	2.4	11.0	0.6
고용노동부	4,178	3.7	19.7	0.1	10.4	6.9	35.2	4.7	10.7	0.3
농림축산식품부	3,511	3.8	19.7	0.0	10.4	6.7	130.1	7.2	12.9	2.5
산업통상자원부	3,403	3.7	28.7	0.6	11.2	5.7	20.6	0.9	10.9	-0.4
과학기술정보통신부	3,161	3.8	16.8	0.1	10.1	5.8	14.3	0.5	9.4	-0.8
국방부	2,995	3.8	19.7	0.2	10.3	6.4	24.5	1.6	11.1	0.8
식품의약품안전처	1,899	3.9	18.5	0.0	10.9	7.7	13.7	0.2	10.4	-0.5
문화체육관광부	1,828	3.7	19.8	0.0	10.2	6.0	22.1	1.0	9.9	-0.3
환경부	1,801	3.7	19.6	0.1	10.3	5.9	28.4	2.7	11.2	0.9
방송통신위원회	1,494	3.9	15.5	0.2	9.4	6.5	44.6	4.7	10.8	1.4
법무부	1,328	4.3	16.7	0.3	9.8	7.0	14.2	0.4	10.1	0.3
인사혁신처	1,027	4.2	17.1	0.2	10.0	5.0	14.2	0.1	10.3	0.3
중소벤처기업부	797	3.7	19.6	0.4	9.8	7.9	48.2	2.4	15.5	5.7
공정거래위원회	774	3.8	19.9	0.0	10.4	7.0	23.5	1.8	11.3	0.9
금융위원회	330	3.9	14.5	0.1	9.0	3.9	28.9	1.2	12.2	3.2
국가보훈처	301	5.4	16.7	0.1	10.5	8.1	12.3	-0.4	10.6	0.1
보건복지부	238	4.0	17.9	0.3	9.9	7.7	13.0	0.3	10.4	0.5
교육부	185	4.0	13.7	-0.2	9.0	6.3	17.6	1.0	9.9	1.0
국토교통부	162	3.7	14.6	0.1	8.6	6.6	15.5	0.6	10.1	1.5
법제처	149	3.9	18.5	0.2	10.7	7.3	23.0	2.6	11.0	0.3
외교부	142	3.8	14.7	0.5	8.8	5.9	14.2	1.0	9.4	0.6
한국산업인력공단	136	4.1	12.8	0.3	7.7	5.0	15.6	0.5	9.6	1.9
통일부	105	5.8	15.7	0.2	9.9	5.8	17.2	1.3	9.6	-0.3

V. 결론

질의응답은 지식과 정보 공유를 위한 대표적 수단이며, 자연어 처리를 비롯하여 다양한 분야에서 오래된 연구 주제이다. 전문가 중심의 답변이 제공되는 CQA 서비스는 대중적 인기를 누리고 있으며 축적된 데이터는 기계학습 분야의 중요한 연구 대상이다. 본 연구에서는 CQA 서비스 중 국민신문고의 민원상담 자료에 대하여 질문과 답변에서 사용된 어휘의 수준별 출현 분포를 시각화하고 해석하였다. 또한, 질의응답에서 사용된 문장의 길이와 어려운 어휘의 비율을 구한 후 Dale-Chall 이독성 공식을 적용하여 문서 수준을 비교하였다. 분석 결과, 질문에서보다 답변에서 어려운 어휘가 많이 사용되고 있음을 확인할 수 있었다. 질문 내용은 민원 신청 방법 및 절차에 관한 것인 반면 답변은 절차 및 방법에 대한 설명이나 제도에 대한 내용이 많아 사용된 단어의 수가 많고 문장의 길이도 길게 나타났다. 어휘 요인과 문장 요인을 반영하여 이독성 지수를 산출했을 때 평균적으로 0.8점, 대략 두 학년의 차이가 있었다. 이독성에 영향을 미치는 요인은 어휘 요인보다는 문장 요인이었으며 이는 문장의 길이를 짧고 간결하게 작성하고 기존 자료를 수정 없이 사용하기 보다는 질문 상황에 맞게 수정하여 답변을 작성할 필요가 있음을 시사한다. 질문에 비하여 답변의 경우 정상 범주를 벗어나는 이독성 지수가 큰 답변이 빈번하게 출현하였는데, 이는 온라인 민원상담시 개선의 여지가 있는 답변을 효율적으로 찾는데 활용할 수 있을 것으로 판단된다.

본 연구에서는 국내의 CQA 서비스에서 제공되는 문서를 대상으로 적합한 이독성 공식과 활용 방안을 제시하고 질의응답 간 이독성 비교를 통하여 상담 개선을 위한 업무적 활용 방안을 모색해 보았다. 작성된 글의 이독성을 자동으로 나타내고 원하는 수준으로 수정할 수 있는 지능적인 환경이 개발된다면 이독성 분석의 활용은 극대화될 것이다. 그러나 본 연구의 핵심요소인 이독성 지수의 정확성은 형태소 분석기의 정확성과 이독성 공식의 정확성 및 타당성에 의존적이라는 한계가 있다. 또한, 이독성 지수의 비교를 위하여 문서를 작성

한 기관으로 그 범위를 제한하였다. 향후 질문 주제, 작성 시기 등 다양한 관점에서 질문과 답변의 이독성을 비교한다면 상담업무 개선을 위한 방안을 마련하는데 더욱 효율적인 정보를 제공할 수 있을 것으로 기대한다.

참 고 문 헌

[1] L. T. Le, C. Shah, and E. Choi, "Evaluating the quality of educational answers in community question-answering," In Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries, pp.129-138, 2016.

[2] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community QA," In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp.411-418, 2010.

[3] https://kin.naver.com/people/partner/, 2019. 10.06.

[4] E. Agichtein, Y. Liu, and J. Bian, "Modeling information-seeker satisfaction in community question answering," ACM Transactions on Knowledge Discovery from Data (TKDD), Vol.3, No.2, Article No.10, 2009.

[5] G. Burel, P. Mulholland, Y. He, and H. Alani, "Predicting answering behaviour in online question answering communities," In Proceedings of the 26th ACM Conference on Hypertext and Social Media, pp.201-210, 2015.

[6] H. Fu and S. Oh, "Quality assessment of answers with user-identified criteria and data-driven features in social Q&A," Information Processing and Management, Vol.56, No.1, pp.14-28, 2019.

[7] J. Trienes and K. Balog, "Identifying Unclear Questions in Community Question Answering Websites," In European Conference on Information Retrieerence on Information Retrieval, pp.276-289, 2019.

[8] https://krdict.korean.go.kr, 2020.01.08.

[9] 최수환, "토픽 모델링을 이용한 사운드스케이프 연구 주제어 분석," 한국콘텐츠학회논문지, 제19권, 제7호, pp.427-435, 2019.

[10] 이재은, 채충일, "기업근로자 경력성공 인식의 다차원성과 차이: 토픽모델링의 적용," 한국콘텐츠학회논문지, 제19권, 제6호, pp.58-71, 2019.

[11] 최정균, 진서훈, 최종후, "텍스트 마이닝 기법을 이용한 언론사별 보도 변화 양상 탐구," J. of the Korean Data Analysis Society, 제19권, 제5호, pp.2509-2522, 2017.

[12] K. Collins-Thompson, "Computational assessment of text readability: A survey of current and future research," J. of Applied Linguistics, Vol.165, No.2, pp.97-135, 2014.

[13] 조용구, "글의 수준을 평가하는 국어 이독성 공식," 독서연구, 제41권, 제3호, pp.73-92, 2016.

[14] 고승연, "이독성 공식 개발을 위한 어휘 목록과 문법 항목의 기대 요인 선정," 한국어문화교육, 제11권, 제2호, pp.1-24, 2018.

[15] W. H. DuBay, *The principles of readability*, Impact Information, 2004.

[16] S. Zhou, H. Jeong, and P. A. Green, "How consistent are the best-known readability equations in estimating the readability of design standards?," IEEE Transactions on Professional Communication, Vol.60, No.1, pp.97-111, 2017.

[17] 윤은정, 박윤배, "과학용어에 대한 '포털 사전', '표준국어대사전', '과학교과서' 설명의 비교 분석," 한국과학교육학회지, 제37권, 제1호, pp.1-8, 2017.

[18] 박정진, "이독성을 활용한 한국어 읽기 자료의 수준 설정 가능성 탐색," 현대사회와 다문화, 제8권, 제2호, pp.1-22, 2018.

[19] 한송화, *한국어 교육 어휘 내용 개발(4단계)*, 국립국어원, 2015.

[20] http://kkma.snu.ac.kr, 2019.10.21.

[21] 정대영, "소설 텍스트의 문장 복잡도 연구 - 자동화된 프로그램을 활용하여 -," 문학교육학, 제48호, pp.236-292, 2015.

저 자 소 개

문 길 성(Gil-Seong Mun) 정회원

- 2001년 2월 : 전북대학교 통계학과(이학사)
- 2003년 2월 : 전북대학교 통계정보학과(이학석사)
- 2011년 2월 : 전북대학교 컴퓨터통계정보학과(이학박사)
- 2004년 3월 ~ 2017년 2월 : 전북대학교 통계학과 강의전담교수
- 2017년 7월 ~ 현재 : 국민연금공단 정보전략실 전문위원

〈관심분야〉 : 기계학습, 사회연결망분석, 통계계산