「정보시스템연구」 제28권 제4호
한국정보시스템학회
2019년 12월, pp. 155～174
http://dx.doi.org/10.5859/KAIS.2019.28.4.155

재정정보 활용을 위한 텍스트 마이닝 기반 회계용어 형태소 분석기 구축*,**

정건용*** · 윤승식**** · 강주영*****

〈목 차〉

Ⅰ. 서 론

재정이란 정부의 재원조달 및 지출활동을 말한다. 기획재정부 열린재정 재정배움터[1]에 의하면 재정은 자원배분의 조정, 소득의 재분배, 경제 안정화 세 가지의 기능을 가진다. 재정정보를 활용한 재정정책은 개인과 기업 모두에게 영향을 미친다. Pérez et al.(2005)은 재정정보를 공개하는 것이 필요하다고 말하며 경제사회 생활에 정보통신기술(ICT)가 일반화되면서 정부차원에서의 정보통신기술 사용 필요성도 늘어났으며, 정부의 재정정보 사용을 투명하게 공개함으로써 국가 및 개인의 경제적 행복에 기여할 수 있다고 주장한다.

* 제1회 '열린재정'을 활용한 대학(원)생 논문 공모전에 수상한 논문임.
** 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터지원사업의 연구결과로 수행되었음.(IITP-2019-2018-0-01424)
*** 아주대학교 e-business학과, munuan00@gmail.com
**** 아주대학교 e-business학과, yoon9577@ajou.ac.kr
***** 아주대학교 e-business학과, jykang@ajou.ac.kr(교신저자)
1) http://www.openfiscaldata.go.kr/portal/baeoom/baeoom01.do(2019.09.30.)

재정정보에는 많은 회계용어들이 사용된다. 국자재정운용의 기초자료를 제공하는 재정정보원의 기획보고서, 분석보고서, 나라재정, 재정통계 등의 자료를 보면 기업들이 사용하는 회계용어부터 정부가 특별기금에 이르기까지 다양한 회계용어가 사용된다. 따라서 재정정보를 분석하기 위해 회계용어에 대한 이해가 필요하다.

정부의 경제활동인 재정정보를 공개하는 것처럼 기업의 경제활동은 재무제표를 통해 공시한다. 재무제표란 회계의 기본목적인 정보이용자에게 유용한 정보를 제공함에 있다. 재무제표는 회사의 현재 상태를 나타내는 대표적인 상태이며 재무적인 위험을 포함한 다양한 정보가 담겨있다(William, 1966). 정부의 특별회계기금 용어와 같은 특별한 경우를 제외하고 정부와 기업에서 쓰는 회계 용어들이 다르지 않기 때문에 정부와 기업에서 공개하는 문서 중 회계용어가 사용된 문서들을 활용하여 재정정보 활용을 위한 자료로 사용할 수 있다. 본 연구에서는 재정정보의 중요성에 주목하여 재정정보 분석을 위한 방법으로 텍스트 마이닝을 제안하려 한다. 텍스트 마이닝을 활용하여 재정정보 및 공공부문 회계에 대해 폭넓게 이해할 수 있다(이원희, 2019).

해외에서는 회계용어가 사용된 텍스트를 활용한 연구가 시도되고 있다. Ravisankar et al.(2011)은 텍스트 마이닝 기법을 활용해 재무제표 텍스트를 분석하여 분식회계를 예측하고 이를 통해 감사인의 업무량을 줄이고자 했다. 국내 또한 재무제표 데이터의 텍스트 마이닝과 관련된 연구가 이루어지고 있다. 텍스트 마이닝 기법을 사용하여 재무제표 주석 내용을 분석해 계량화함으로써 재무제표의 정보 비대칭 수준을 낮추는 연구가 이루어지고 있다(모예린 등, 2019).

하지만 한국어는 텍스트 마이닝에 선행되는 형태소 분석에 있어 몇 가지 어려움이 있다. 엄격하게 형태론적 처리를 한다면 실제 형태소 사전에 없는 후보군들이 생성되며 분석 속도를 저하시키는 원인으로 작용한다(심광섭, 2013). 또한 어미에 변화에 따라 구축한 모델의 성능이 다르게 나타나기도 한다(김정호 등, 2010). 게다가 재무제표에 등장하는 단어들은 일반적인 단어가 아닌 회계분야의 전문용어들이 사용되어 일반적인 형태소 분석기로 텍스트를 분석했을 때 좋은 결과를 기대하기 어렵다.

따라서 본 연구에서는 재정정보의 특성을 반영하여 회계용어가 사용된 텍스트를 분석하기 위한 형태소 분석기를 구축할 것이다. 재무제표에는 다양한 회계용어가 사용되고 있으며 적절한 형태소 분석기 구축을 위해 재무제표 데이터를 사용할 것이다. 재무제표에서 가장 많은 텍스트 정보를 포함하는 주석, 재정정보원의 공시자료, 삼일회계법인의 재정과목 마스터를 기반으로 재정정보의 특성을 담은 형태소 분석기를 구축할 것이다. 구축 방법으로는 재정정보원의 회계·기금운용구조에 있는 텍스트 자료와 국내 전기전자업 코스피 시가총액 상위 10개 기업이 발행한 분기, 반기, 사업보고서와 기재정정 보고서의 주석 텍스트를 활용할 것이다. 텍스트에서 명사를 추출한 뒤 N-gram 기법을 활용해 특징 말뭉치(corpus)를 추출하고 이를 활용해 상용 형태소 분석기를 튜닝할 것이다. 이 때, 삼일회계법인 K-IFRS계정과목 마스터를 추가로 활용하여 분석기의 정확도를 높일

것이다. 이를 통해 분석된 재무제표 텍스트는 기업의 재무제표를 분석하는 데에 유용하다는 장점이 있을 것이다. 전기전자업을 데이터 수집 대상으로 선정한 것은 단순 편의로 산업군을 나눈 것일 뿐이며 분석하길 원하는 산업군 혹인 기업들의 재무제표 주석 데이터를 활용한다면 해당 분야의 특성이 반영된 형태소 분석기를 구축할 수 있을 것이다.

본 연구의 구성은 다음과 같다. 제 2장에서 해외에서 이루어진 재무제표 텍스트 분석과 국내의 재무제표 텍스트 분석에 대한 연구를 작성하였다. 그리고 특정 산업을 위한 텍스트 분석기 구축에 관한 연구를 정리하였다. 마지막으로 본 연구에서 사용될 텍스트 분석 기법에 대해 기술하였다. 제 3장은 수집한 3가지 데이터(재무제표 주석, 재정정보원 회계·기금운용, 삼일회계법인 K-IFRS 계정과목 마스터)를 기반으로 텍스트 마이닝 기법을 활용한 회계용어 형태소 분석기 구축 방법을 소개하였다. 제 4장에서는 연구방법에 대한 소개와 실제 재무제표에 사용된 문장을 통해 새로운 형태소 분석기의 성능을 증명하였다. 제 5장에서는 분석 결과의 시사점과 추후 연구방향 그리고 한계점에 대해 작성하였다.

Ⅱ. 선행연구

2.1 해외 및 국내 재무제표 텍스트 분석

국내, 외에서 재무제표의 유용성을 높이기 위한 다양한 연구가 이루어지고 있다. 유용성을 높이기 위한 방법 중 하나로 텍스트 마이닝이 사용되고 있으며 텍스트 마이닝을 통해 재무제표에 숨겨진 내용을 밝혀내기 위한 연구가 이루어지고 있다. Kamaruddin et al.(2015)은 재무제표 데이터를 기반으로 텍스트 마이닝 기법을 사용할 시 문서 내에 존재하는 비정상 문장을 탐지할 수 있으며 이를 위해 재무제표 내의 텍스트 데이터와 수치 데이터를 수집하여 사용할 수 있다고 한다.

현재까지의 연구는 회계처리기준을 위반한 기업을 예측하기 위한 연구가 활발하다. 재무제표 텍스트 분석을 통해 재무제표 재작성 가능성이 높은 기업을 사전에 예측하여 재무제표 재작성을 사전에 방지하기 위한 연구이다. 박경진 등(2014)은 재무제표 재작성은 외부이용자에게 혼란을 유발할 수 있다고 한다. 재무제표 재작성을 유발하는 행위 중 한가지로 오류수정 효과를 당기손익에 반영함으로써 당기순이익이 왜곡되고, 당기손익에 반영되어야 할 항목이 오류수정을 통한 재작성의 형태로 공시되기도 한다고 말하며 이는 자본시장의 효율적인 기능을 방해하며 투자자의 손실을 초래할 수 있다. 다음은 재무제표 재작성을 유발하는 2가지 요인에 관한 선행연구이다.

첫 번째로 재무제표 재작성을 유발하는 고의적인 회계처리기준 위반과 관련된 내용이다. DECHOW et al.(2011)은 수익을 인식하지 않고 조작하는 기업들이 많이 존재한다고 말한다. 고의적인 분식회계 기업을 예측하기 위해 재무제표 재작성 사례를 포함한 포괄적인 데이터셋 구축이 필요하며, Ravisankar et al.(2011)은 축적된 재무제표 데이터를 기반으로 다양한 텍스트 마이닝 기법을 적용하여 분식회계를 예측할

수 있다고 주장한다.

두 번째로 재무제표 재작성을 유발하는 단순 오류로 인한 회계처리 기준 위반과 관련된 내용이다. 분식회계로 인한 재작성을 포함하여 단순 오류로 인한 재작성을 고려한 연구도 있다. Dutta et al.(2017)은 단순 오류로 인한 재작성도 자본시장의 효율성에 해로울 수 있으며 투자자들의 신뢰를 잠식할 수 있다고 한다. 단순 오류로 인한 재작성을 관리하지 않는다면 내부통제 환경이 약화되고 경영상의 감독이나 잘못된 보고를 억제하기 위한 노력이 줄어들 수 있다. 고의적인 분식회계 사례와 단순오류로 인한 재작성 사례를 모두 포함한 데이터셋을 사용하여 여러 가지 텍스트 마이닝 기법(의사결정 나무, 서포트 벡터 머신, 베이지안 네트워크)등을 통해 재무제표 재작성을 예측하는 유용한 모델을 개발할 수 있다고 말한다.

2.2 특정 분야를 위한 형태소 분석기 구축

이현영 등(2016)은 한국어는 특성상 생략이 자주 일어나고 수식이 자유로우며 도치의 사용, 접사의 사용 등 언어학적 특성 때문에 파싱을 하는 것보다 형태소 분석 단계에서 처리하면 좋을 수 있다고 말한다. 하지만 모든 경우에 통용되는 범용 형태소 분석기를 구축하는 것은 쉽지 않다. 각 분야에 사용되는 단어의 형태가 다르며 전문용어가 사용되고 그 분야에서 새로 생성되는 단어가 많기 때문이다.

모든 규칙을 일일이 처리하는 것은 처리속도를 드리게 하거나 메모리 사용 범위에 부담을 줄 수 있다. 송은지(2015)에 의하면 모든 단어 사용 규칙을 일일이 처리하는 것은 시스템의 효율성을 따지는 컴퓨터 프로그램 분야에서 좋은 해결법이 아니라고 한다. 넓은 보편적인 언어 사용 현상을 규칙화 하는 것이 좋지만, 사용 범위가 좁은 단어들에 대해서는 예외적인 규칙을 사용해야 한다. 따라서 일반적인 범용 형태소 분석기에 각 분야별 특성화된 규칙을 추가할 필요가 있다.

일반적인 문서를 기반으로 구축된 기존 형태소 분석기를 특정 분야에 활용하였을 때 좋은 성능을 기대하기 어려운 경우가 있다. 강승식(2008)은 SMS영역에 일반적인 형태소 분석기를 사용하면 등록되지 않은 언어에 대한 분석 오류가 발생하는 경우가 많으며 복합명사를 분해하는 과정에서 문제가 일어날 수 있다고 한다. 일반적인 형태소 분석기를 경제 분야에 사용했을 때에도 좋은 성능을 기대하기 어렵다는 문제를 해결하기 위해 경제 분야에 최적화된 사전을 구축하기 위한 연구가 있다. 박기영 등(2019)은 금융통화위원회 의사록 분석을 위한 경제분야의 용어사전을 사용하였으며 형태소 분석기를 수정한 eKoNLPy를 제작하였다. 사전 구축을 위해 동의어, 외래어처리와 N-gram 기법을 사용한 사전 구축 방법을 사용하였다. 이정민 등(2016)은 무용학에 관한 자료를 텍스트 마이닝으로 처리하는 기초 마련을 위해 연구를 수행했다. 무용 관련 저널의 데이터를 기반으로 형태소 분석을 실시한 뒤 분석 내용을 기반으로 새로운 사용자 사전을 추가한 형태소 분석기를 구축하였다. 작품, 장르, 이론 분야의 명사형 단어를 사전에 추가하여 향상된 결과를 얻을 수 있었다.

2.3 N-gram

N-gram이란 통계적 언어모델(Statistical Language Model)의 일종이다. N은 연속적인 단어 나열의 개수를 의미한다. N개의 연속적인 단어 나열을 활용하여 그 다음에 올 단어의 확률을 예측하는데 사용된다. N-gram 기법은 문장 사이에 숨겨진 문맥을 파악할 수 있게 한다. 한 가지 단어가 여러 의미를 가질 수 있기 때문에 단어 하나로는 문맥을 고려하기 힘든 경우가 있다. 예를 들면 '교정(校庭)'은 학교 운동장의 의미를 갖기도 하지만 잘못된 것을 바로잡다 는 의미의 '교정-하다'의 원형으로 쓰이기도 한다. 한 개 이상의 단어가 연속적으로 사용될 때 긍, 부정의 어조가 바뀌기도 한다. 예를 들면, 'unemployment'는 실업률을 의미하는데 'lower unemployment'는 낮은 실업률을 의미한다(이준영 등, 2019). 따라서 하나의 단어보다 여러 가지 단어를 묶어서 분석하는 것이 문맥을 더 잘 파악할 수 있다.

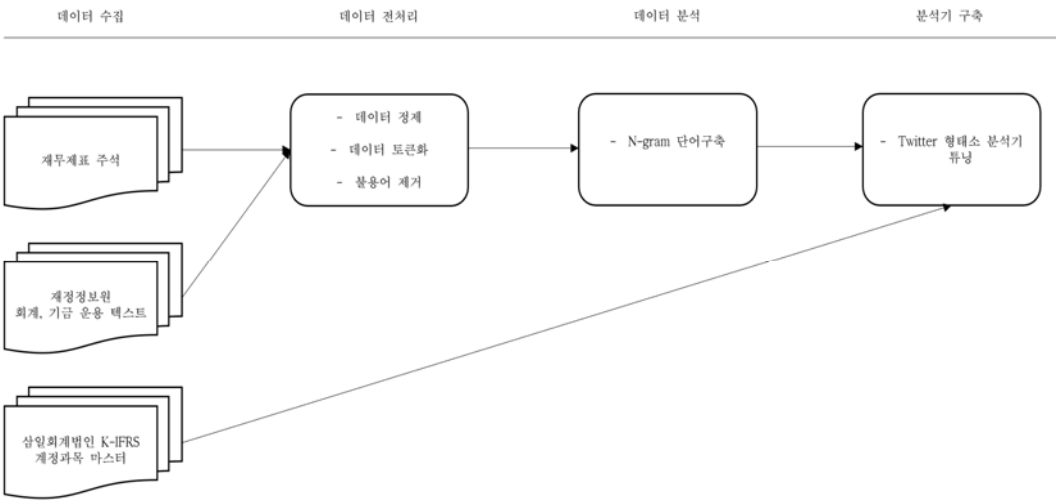
위 장점 때문에 N-gram은 텍스트 마이닝을 위한 전단계인 사전 구축을 위해 많이 사용된다. 권희준 등(2012)은 악성코드 분류 시스템 구축을 위해 N-gram을 사용한다. 컴퓨터 언어를 N-gram 방법으로 추출해서 Matrix 생성 후 악성코드 분류에 사용한 결과 정확도를 높일 수 있었다고 한다. Tripathy et al. (2016)은 감성분석에 N-gram을 활용한 말뭉치(corpus)를 활용하였으며 Uni-gram을 사용할 때 보다 Tri-gram, 4-gram, 5-gram을 함께 사용할 때 모델의 성능이 올라갔다고 한다.

Ⅲ. 연구방법

3.1 연구 주제

본 연구에서는 주석 텍스트와 재정정보원의 회계보고서에 있는 텍스트와 삼일회계법인 K-IFRS계정과목 마스터를 활용하여 재무제표의 특성을 담은 형태소 분석기를 구축하고자 한다. 황선필 등(2017)은 (K-GAAP)에 비해 한국채택국제회계기준(K-IFRS)에 의해 공시되는 정보는 양적으로 증가했다고 긍정적으로 보지만, 상투적인 표현과 어려운 정보 등으로 인해 핵심정보 파악이 힘들다고 평가했다. 또한 재무제표에 사용되는 단어는 하나의 명사로 이루어진 형태뿐만 아니라 명사-명사, 명사-접사 등의 형태를 띠고 있다(장진영, 2012). 기본적인 텍스트 마이닝 접근 방법은 문서에서 단어를 추출하여 단어-문서 매트릭스를 구성하는 것이다(Perkins and Jacob, 2014). 이때 단어 추출 방법에서 사용되는 형태소 분석기에 일반적인 형태소 분석기를 사용한다면, 세부감사계획과 같은 단어를 세부, 감사, 계획으로 나누기 때문에 재무제표의 단어를 효과적으로 검출할 수 없다.

본 연구에서는 이러한 문제를 해결하기 위해 텍스트 마이닝 기법을 활용하여 재무제표의 특성 단어를 구분하고자 했다. 텍스트 마이닝 기법 중 N-gram은 문장의 희소성 문제를 완화하기 위한 대표적 방법으로 문장 내 존재하는 단어를 Uni-gram뿐만 아니라 N-gram까지 확장하여 사용하는 방법이다(Post and Bergsma, 2013). 따라서 텍스트 마이닝 기법인 N-gram을 활용하여, 재무제표의 특성 있는 단어를 구분하



<그림 1> 연구 프레임워크

고 완성도 높은 형태소 분석기를 구축할 수 있을 것으로 기대하였다.

3.2 연구 방법

앞에서 제시한 연구주제를 위한 본 연구의 과정은 <그림 1>과 같이 데이터를 수집하고 분석을 위한 전처리를 거친 뒤, 텍스트 마이닝 기법인 N-gram과 이를 통해 도출된 단어의 조합을 twitter 형태소 분석기에 추가하여 재무제표 형태소 분석기를 구축할 것이다.

수집할 데이터는 다음 3가지이다. 첫 번째 재무제표의 주석데이터. 두 번째 재정정보원의 회계·기금운용구조 특별회계관련 텍스트 데이터, 마지막으로 삼일회계법인이 제공하는 K-IFRS 계정과목 마스터 데이터이다. 3가지 데이터를 이용하여 형태소 분석기를 구축하려는 이유는 다음과 같다. 장진영 (2012)에 의하면 우리나라에서 사용되는 회계용어는 새로운 회계용어를 만들어 사용하거나 국제회계기준(IFRS)를 번역하여 사용한다. 새로 만들어진 회계용어를 대비하기 위하여 회사 내의 재무제표 주석 데이터를 사용했으며 국제회계기준을 번역하여 사용하는 것을 대비하기 위하여 삼일회계법인의 K-IFRS 계정과목 마스터 데이터를 사용하였다.

첫 번째, 재무제표 주석은 재무제표 본문에 기록된 내용에 더하여 정보이용자가 본문의 내용을 더 잘 이해할 수 있게 돕는 역할을 수행한다. 주석은 재무제표의 구성 요소 중 가장 많은 텍스트 정보가 존재하며 재무제표 본문의 내용을 부연설명 하거나 인식 요건을 충족하지 못한 내용이 기록되어 재무제표 이용자에게 도움을 줄 수 있다. 즉 주석이 없다면 재무제표의 내용을 충분히 이해할 수 없다. 두 번째, 재정정보원의 회계·기금운용구조 특별회계관련 텍스트 데이터에는 일반적으로 재무제표 주석 작

성에 사용되지 않는 용어들이 담겨있다. 해당 특별회계의 근거법과 세입, 세출에 관한 전문 용어들이 사용되고 있다. 새로 구축한 형태소 분석기가 특별회계와 관련된 용어를 인식하지 못한다면 재무제표의 텍스트를 적절히 분석하지 못할 것이다. 세 번째, 삼일회계 법인이 제공하는 K-IFRS 계정과목 마스터 데이터는 “한국채택국제회계기준과 일반기업회계기준에 대한 이해를 돕기 위해 기준서 해설자료 및 회계처리, 질의회신 및 기타 기고문, 공시사항점검표 및 주석사례 등의 관련 자료를 사용자의 편의를 위해 계정과목별로 분류한 것”이라고 한다.[2] 이 분류표에 나오는 계정과목들은 실제 재무제표를 작성할 때 빈번하게 등장하는 단어들이며 이 단어를 사전에 추가하는 것은 재무제표의 텍스트를 적절히 분석할 수 있게 할 것이다.

회계용어 분석에 특화된 형태소 분석기 구축을 위해 수집한 3가지의 데이터를 2가지 방법으로 활용 할 것이다. 먼저 첫 번째 데이터인 재무제표의 주석데이터와 두 번째 데이터인 재정정보원의 회계·기금운용구조 특별회계관련 텍스트 데이터는 N-gram 기법을 활용할 것이다. 두 가지 데이터를 토큰화 하여 명사를 추출해낸 후 N-gram 기법을 활용하여 생성한 새로운 말뭉치(corpus)를 형태소 분석기에 추가 할 것이다. 세 번째 데이터인 삼일회계법인이 제공하는 K-IFRS 계정과목 마스터 데이터는 실제 재무제표에서 사용되는 단어이기 때문에 단어를 그대로 형태소 분석기에 추가 할 것이다.

Ⅳ. 연구 분석 및 결과

4.1 데이터 수집

형태소 분석기 구축에 사용된 데이터는 DART에 공시된 재무제표 주석데이터, 재정정보원 2019 회계・기금운용구조 특별회계관련 텍스트 데이터, 삼일회계법인 K-IFRS 계정과목 마스터 데이터이다. 먼저 DART(전자공시시스템)에서 제공하는 Open API를 활용하여 재무제표를 크롤링 하였다. DART Open API는 DART 홈페이지에서 제공하는 공시정보를 외부에서 이용가능하게 한 서비스이다. 최근 10년 동안의 재무제표 주석정보를 Python을 활용해 수집했다. 발급받은 API Key와 요청 주소, 요청 변수들을 활용해 데이터셋을 구축하였다. 검색 시작일은 10년 전(2009년)으로, 공시 성격은 분기, 반기, 사업보고서와 기재 정정된 사업보고서의 주석내용을 수집하였다. 수집 대상은 산업군 별 단어 구축을 위한 시작으로 한국거래소(KRX) KRX 지수산업분류의 코스피 산업지수 중 전기전자 산업 상위 시가총액 10개기업 중 삼성전자(우)를 제외한 9개 기업을 수집하였다. 수집한 데이터의 메타데이터는 다음 <표 1>과 같다.

다음 <표 2>는 크롤링을 통해 수집한 데이터의 일부이다. 법인 구분, 법인 명, 종목코드, 공시구분, 접수번호, 공시 제출인명, 접수일자, 주석 이다.

두 번째로 재정정보원 2019 회계・기금운용구조의 특별회계 목적으로 공시된 총 19항목의 데이터(교도작업특별회계, 교통시설특별회계,

2) https://www.samili.com/acc/kijun/Kifrs_tot_list.asp?op=1&op2=1(2019.09.30.)

<표 1> 전자공시시스템(DART) Open API 요청변수

구분	내용
crp_cls	법인구분 : Y(유가), K(코스닥), N(코넥스), E(기타)
crp_nm	공시대상회사의 종목명(상장사) 또는 법인명(기타법인)
crp_cd	공시대상회사의 종목코드(6자리) 또는 고유번호(8자리)
rpt_nm	공시구분+보고서명+기타정보 [기재정정] : 본 보고서명으로 이미 제출된 보고서의 기재내용이 변경되어 제출된 것임 [첨부정정] : 본 보고서명으로 이미 제출된 보고서의 첨부내용이 변경되어 제출된 것임 [첨부추가] : 본 보고서명으로 이미 제출된 보고서의 첨부서류가 추가되어 제출된 것임 [변경등록] : 본 보고서명으로 이미 제출된 보고서의 유동화계획이 변경되어 제출된 것임 [연장결정] : 본 보고서명으로 이미 제출된 보고서의 신탁계약이 연장되어 제출된 것임 [발행조건확정] : 본 보고서명으로 이미 제출된 보고서의 유가증권 발행조건이 확정되어 제출된 것임 [정정명령부과] : 본 보고서에 대하여 금융감독원이 정정명령을 부과한 것임 [정정제출요구] : 본 보고서에 대하여 금융감독원이 정정제출 요구를 부과한 것임
rcp_no	접수번호(공시뷰어 연결에 이용) - PC용 : http://dart.fss.or.kr/dsaf001/main.do?rcpNo=접수번호 - 모바일용 : http://m.dart.fss.or.kr/html_mdart/MD1007.html?rcpNo=접수번호
flr_nm	공시 제출인명
rcp_dt	공시 접수일자(YYYYMMDD)
rmk	조합된 문자로 각각은 아래와 같은 의미가 있음 유 : 본 공시사항은 한국거래소 유가증권시장본부 소관임 코 : 본 공시사항은 한국거래소 코스닥시장본부 소관임 채 : 본 문서는 한국거래소 채권상장법인 공시사항임 넥 : 본 문서는 한국거래소 코넥스 시장 소관임 공 : 본 공시사항은 공정거래위원회 소관임 연 : 본 보고서는 연결부분을 포함한 것임 정 : 본 보고서 제출 후 정정신고가 있으니 관련 보고서를 참조하시기 바람 철 : 본 보고서는 철회(간주)되었으니 관련 철회신고서(철회간주안내)를 참고하시기 바람
note	재무제표 주석

자료 : 전자공시시스템(http://dart.fss.or.kr/dsap001/guide.do)

<표 2> 수집된 재무제표 주석 데이터

crp_cls	crp_nm	crp_cd	rpt_nm	rcp_no	flr_nm	rcp_dt	note
Y	삼성SDI	006400	반기보고서	20190814001473	삼성SDI	20190814	해당주석
Y	삼성SDI	006400	분기보고서	20190515001002	삼성SDI	20190515	해당주석
Y	삼성SDI	006400	사업보고서	20190329003979	삼성SDI	20190329	해당주석
Y	삼성SDI	006400	분기보고서	20181114001947	삼성SDI	20181114	해당주석

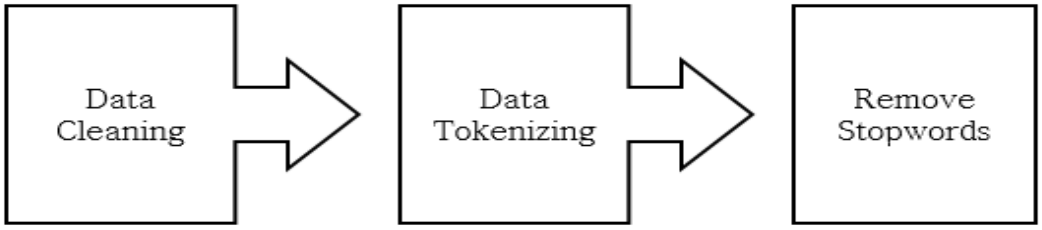
국가균형발전특별회계, 국방·군사시설이전특별회계, 농어촌구조개선특별회계, 등기특별회계, 아시아문화중심도시조성특별회계, 에너지 및 자원사업특별회계. 우체국보험특별회계. 유아교육지원특별회계, 주한미군기지이전특별회계, 행정중심복합도시건설특별회계, 혁신도시건설특별회계, 환경개선특별회계, 양곡관리특별회계, 우체국예금특별회계, 우편사업특별회계, 조달특별회계, 책임운영기관특별회계)와 관련된 모든 텍스트 데이터를 활용하였다. 다음 <표 2>는 재정정보원의 데이터 중 특별회계 관련 데이터이다.

세 번째로 삼일회계법인의 계정과목 마스터 데이터를 활용하였다. 계정과목 마스터는 회계처리의 편의와 주석사례 등 이용자의 편의를 위해 참고목적으로 분류해 둔 결과이다. IFRS의 모든 내용을 담고 있지는 않지만 일반적으로 자주 사용되는 단어들임으로 형태소 분석기 구축에 사용하였다. 다음 <표 3>은 삼일회계법인에서 제공하는 K-IFRS 계정과목 마스터 데이터이다.

<표 3> 삼일회계법인 K-IFRS 계정과목 마스터

구분		계정과목등		
재무제표		재무상태표	포괄손익계산서	자본변동표
		현금흐름표	연결재무제표	별도재무제표
		주석 및 부속명세서		
자산	금융자산	현금 및 현금성자산	당기손익인식금융자산	장단기대여금
		장단기매출채권	대손충당금	기타수취채권
		매도가능금융자산	만기보유금융자산	
	비금융자산	선급비용 및 선급금	재고자산	투자부동산
		유형자산	무형자산	관계기업/공동기업
		이연법인세자산		
부채	금융부채	장단기차입금 및 사채	전환사채	장단기매입채무
		미지급비용	미지급금	신주인수권부사채
		금융보증계약	보증금	
	비금융부채	선수금	예수금	선수수익
		퇴직급여부채	충당부채	우발부채
		이연법인세부채	보고기간후사건	
자본		납입자본	자기주식	자기주식처분손익
		감자차손익	주식선택권	기타자본
		기타포괄손익누계	이익잉여금	
손익항목		매출액	매출원가	판매비와 관리비
		금융수익원가	기타수익비용	법인세비용
		주당이익		
기타		중간재무보고	매각예정자산	중단영업
		회계변경과 오류수정	공정가치	
특수회계처리		파생상품	리스	공동약정
		사업결합	한국채택국제회계기준의 최초채택	농림어업

자료 : 삼일회계법인(https://www.samili.com/acc/kijun/Kifrs_tot_list.asp?op=1&op2=1)



<그림 2> 데이터 전처리 흐름도

4.2 데이터 전처리

비정형 데이터인 텍스트를 분석하기 위해서는 데이터를 전처리 하는 것이 필요하다. 데이터 전처리는 다음 <그림 2>와 같이 데이터 정제 - 데이터 토큰화 - 불용어 처리 순서로 진행했다.

삼일회계법인 K-IFRS 계정과목 마스터 데이터는 데이터 전처리 없이 단어 자체를 명사로 등록하였다.

먼저, 데이터 정제 단계에서는 전자공시시스템에서(DART)에서 제공하는 API를 사용하여 크롤링한 재무제표 주석 데이터와 재정정보원 2019 회계·기금운용구조 특별회계관련 텍스트 데이터에 포함된 “\n, <, >,“”, ‘’, -, ~”등의 특수단어를 제거하였다. 기재정정 보고서의 경우 재무제표 주석 데이터가 비어있는 경우가 있었으며 이는 NULL값으로 대체하였다.

두 번째 데이트 토큰화 단계에서는 다음과 같이 전처리 하였다. 장진영 (2012)은 회계용어를 분석한 결과 전체 1029 용례 중 828개의가 명사어 용례로서 80.46%를 차지한다고 분석했다. 명사로 이루어져 있으며 명사와 명사, 접미사, 접두사로 이루어진 경우가 그 뒤를 잇는다고 한다. 따라서 데이터 토큰화를 할 때 명사로 토큰화 하였으며 cKonoly의 twitter tokenizer를 사용하였다.

마지막으로 불용어 처리 과정은 다음과 같이 진행했다. 명사로 토큰화한 리스트 중 불용어를 제거하였다. 장진영 (2012)은 회계용어를 분석한 결과 전체 1029개의 용례 중 993개가 한자어 용례로서 96.5%를 차지한다고 분석했다. 따라서 명사이면서 한자어가 아닌 것과, 단어를 조합했을 때 회계용어가 되기 힘든 단어들을 불용어처리 하였다. N-gram방법 사용 후 사전 구축을 하기 위해 새로운 불용어 처리임으로 다음과 같은 기준으로 처리하였다.

4.3 형태소 분석기 구축

명사로 이루어진 회계용어 말뭉치(corpus)를 형태소분석기 구축에 활용하기 위해 명사로 토큰화된 데이터를 사용하여 n-gram을 적용하였다. 분석기 구축에는 n<=5인 n-gram을 사용하였다. 차원의 저주로 n이 클수록 분석기 구축에

사용된 말뭉치(corpus)에서 해당 n-gram을 발견할 확률이 적어지기 때문에 희소문제가 커질 수 있다. 따라서 n이 5를 넘어가지 않게 하였다.

<표 4> 불용어 처리 기준

제거기준	예
한자어가 아닌 것	시스템, 소프트웨어, 골프, 코퍼레이션, 에누리, 픽스 등
인명	박준, 이성용, 김용균 등
직책	상무이사, 사장, 임원 등
회사/공공기관명	에스케이텔레콤, 에스케이씨앤씨, 외환은행, 한국수출입은행, 구미시청, 한국거래소 등
국가/지역/도시명	유럽, 베트남, 텍사스, 영등포구, 아시아, 서울특별시 등
화폐단위	엔화, 위안화, 달러 등
회계용어가 아닌 명사	거의, 반올림, 이제, 한번, 나 등

<표 5> 형태소 분석기 사용 결과(단어)

형태소 분석기 구축에 사용된 말뭉치(corpus)	
민사소송	과세표준
매매정지	기초자산
내부거래	독점계약
채무불이행	균형발전특별회계
국민연금기금	과학기술정보통신부

<표 6> 형태소 분석기 사용 결과(단어)

기존 형태소 분석기	새로운 형태소 분석기
‘공정’,‘가치’	‘공정가치’
‘법인세’, ‘비용’	‘법인세비용’
‘사용권’, ‘자산’	‘사용권자산’
‘계약’, ‘이행’	‘계약이행’
‘지급’, ‘보증’	‘지급보증’
‘반기’, ‘재무제표’	‘반기재무제표’
‘자본’, ‘관리’, ‘지표’	‘자본관리지표’
‘기업’, ‘고유’, ‘정보’	‘기업고유정보’
‘환율’, ‘변동’, ‘위험’	‘환율변동위험’
‘유동성’, ‘장기’, ‘부채’	‘유동성장기부채’
‘공정’, ‘가치’, ‘금융’, ‘자산’	‘공정가치금융자산’
‘보고’, ‘기간’, ‘종료’, ‘일’	‘보고기간종료일’
‘가중’, ‘평균’, ‘연간’, ‘법인’, ‘세율’	‘가중평균연간법인세율’

형태소 분석기 구축에는 n-gram을 활용하여 n이 5를 넘지 않는 총 190만개의 말뭉치가 사용되었으며 다음은 실제 형태소분석기 구축에 사용된 말뭉치(corpus)의 일부이다.

4.4 형태소 분석기 구축 후 재무제표 텍스트 분석 결과

다음 <표 6>은 삼성전자의 2019년 8월 발행된 반기보고서에 실제 사용된 단어를 사용하여 기존 형태소 분석기와 형태소 분석기 구축 후 새로운 형태소 분석기를 사용한 결과를 비교한 표이다. 회계용어가 여러 개의 분할되어 인식되는 기존의 형태소 분석기와 달리 새로 구축한 형태소 분석기는 회계용어를 인식한다. 기존 형태소 분석기는 cKonlpy의 twitter를 사용하였으며, 새로 구축한 형태소 분석기는 기존 분석기를 기반으로 새로운 단어를 추가하였다.

다음 <표 7>은 삼성전자의 2019년 8월 발행된 반기보고서에서 실제 사용된 문장을 사용하여 기존 형태소 분석기와 형태소 분석기 구축 후 새로운 형태소 분석기를 사용한 결과를비교한 표이다. <표 6>에서 단어단위로 분석한 것과 마찬가지로 분할되어 인식되던 회계용어가 분할 되지 않고 인식되는 것을 확인할 수 있다.

<표 7> 형태소 분석기 사용 결과(문장)

재무제표 문장 중 일부	기존 형태소 분석기	새로운 형태소 분석기
보고기간종료일 현재 다수의 회사 등과 정상적인 영업과정에서 발생한 소송, 분쟁 및 규제기관의 조사 등이 진행 중에 있습니다.	('보고', 'Noun'), ('기간', 'Noun'), ('종료', 'Noun'), ('일', 'Noun'), ('현재', 'Noun'), ('다수', 'Noun'), ('의', 'Josa'), ('회사', 'Noun'), ('등', 'Noun'), ('과', 'Josa'), ('정상', 'Noun'), ('적', 'Suffix'), ('인', 'Josa'), ('영업', 'Noun'), ('과정', 'Noun'), ('에서', 'Josa'), ('발생', 'Noun'), ('한', 'Josa'), ('소송', 'Noun'), (',', 'Punctuation'), ('분쟁', 'Noun'), ('및', 'Noun'), ('규제', 'Noun'), ('기관', 'Noun'),	('보고기간종료일', 'Noun'), ('현재', 'Noun'), ('다수', 'Noun'), ('의', 'Josa'), ('회사', 'Noun'), ('등', 'Noun'), ('과', 'Josa'), ('정', 'Noun'), ('상적', 'Noun'), ('인', 'Josa'), ('영업과정', 'Noun'), ('에서', 'Josa'), ('발생', 'Noun'), ('한', 'Josa'), ('소송', 'Noun'), (',', 'Punctuation'), ('분쟁', 'Noun'), ('및', 'Noun'), ('규제기관', 'Noun'), ('의', 'Josa'), ('조사', 'Noun'), ('등', 'Noun'), ('이', 'Josa'), ('진행', 'Noun'),

	('의', 'Josa'), ('조사', 'Noun'), ('등', 'Noun'), ('이', 'Josa'), ('진행', 'Noun'), ('중', 'Noun'), ('에', 'Josa'), ('있습니다', 'Adjective'), ('.', 'Punctuation')	('중', 'Noun'), ('에', 'Josa'), ('있습니다', 'Adjective'), ('.', 'Punctuation')
법인세비용은 전체 회계연도에 대해서 예상되는 최선의 가중평균연간법인세율의 추정에 기초하여 인식하였습니다. 당반기 현재 2019년 12월 31일로 종료하는 회계연도의 예상평균 연간법인세율은 15.5%입니다.	('법인세', 'Noun'), ('비용', 'Noun'), ('은', 'Josa'), ('전체', 'Noun'), ('회계', 'Noun'), ('연도', 'Noun'), ('에', 'Josa'), ('대해', 'Noun'), ('서', 'Josa'), ('예상', 'Noun'), ('되는', 'Verb'), ('최선', 'Noun'), ('의', 'Josa'), ('가중', 'Noun'), ('평균', 'Noun'), ('연간', 'Noun'), ('법인', 'Noun'), ('세율', 'Noun'), ('의', 'Josa'), ('추정', 'Noun'), ('에', 'Josa'), ('기초', 'Noun'), ('하여', 'Verb'), ('인식', 'Noun'), ('하였습니다', 'Verb'), ('.', 'Punctuation'), ('당', 'Modifier'), ('반기', 'Noun'), ('현재', 'Noun'), ('2019년', 'Number'), ('12월', 'Number'), ('31일', 'Number'), ('로', 'Foreign'), ('종료', 'Noun'), ('하는', 'Verb'), ('회계', 'Noun'), ('연', 'Modifier'), ('도의', 'Noun'),	('법인세비용', 'Noun'), ('은', 'Josa'), ('전체', 'Noun'), ('회계연도', 'Noun'), ('에', 'Josa'), ('대해', 'Noun'), ('서', 'Josa'), ('예상', 'Noun'), ('되는', 'Verb'), ('최선', 'Noun'), ('의', 'Josa'), ('가중평균연간법인세율', 'Noun'), ('의', 'Josa'), ('추정', 'Noun'), ('에', 'Josa'), ('기초', 'Noun'), ('하여', 'Adverb'), ('인식', 'Noun'), ('하였습니다', 'Verb'), ('.', 'Punctuation'), ('당', 'Noun'), ('반기', 'Noun'), ('현재', 'Noun'), ('2019년', 'Number'), ('12월', 'Number'), ('31', 'Number'), ('일로', 'Noun'), ('종료', 'Noun'), ('하는', 'Verb'), ('회계연도', 'Noun'), ('의', 'Josa'), ('예상평균', 'Noun'), ('연간법인세율', 'Noun'), ('은', 'Josa'), ('15.5%', 'Number'), ('입니다', 'Adjective'), ('.', 'Punctuation')

	('예상', 'Noun'), ('평균', 'Noun'), ('연간', 'Noun'), ('법인', 'Noun'), ('세율', 'Noun'), ('은', 'Josa'), ('15.5%', 'Number'), ('입니다', 'Adjective'), ('.', 'Punctuation')	

다음은 삼성전자의 2009년 ~ 2019년간 재무제표 주석데이터를 시각화 한 것이다. 기존의 형태소 분석기로 추출한 명사 단어와 새로운 형태소 분석기로 추출한 명사 단어의 빈도수를 집계하여 워드클라우드를 그렸다. <그림 3>의 단어들은 하나의 명사로만 이루어져 있는 것을 확인할 수 있으며 <그림 4>의 단어들은 명사뿐 아니라 명사구 가 인식되는 것을 확인할 수 있다.

<그림 3> 기존 분석기 워드클라우드

<그림 4> 새로운 분석기 워드클라우드

Ⅴ. 결론 및 시사점

5.1 결과 및 토의

본 연구에서는 재정정보를 활용하여 기존 형태소분석기를 회계용어 분석에 용이하게 튜닝하였다. 기존 형태소 분석기의 문제점을 개선하여 회계용어가 사용된 텍스트를 분석할 때 기존 형태소 보다 목적적합한 용어가 추출되도록 도울 수 있게 하였다. 이를 위해 재무제표 주석데이터, 재정정보원 회계 및 기금 운용구조 특별회계관련 텍스트 데이터, 삼일회계법인

K-IFRS 계정과목 마스터 데이터를 수집하였다. 수집된 데이터는 데이터 정제, 데이터 토큰화, 불용어 제거의 과정을 거쳤다. 전처리된 데이터를 활용해 N-gram 방법에 의해 생성된 단어를 새로운 단어로서 기존 형태소 분석기에 추가하였다. 새로 구축된 형태소 분석기를 활용해 재무제표에 사용된 단어와 문장을 분석하였을 때 나오는 단어를 기존 형태소 분석기를 활용했을 때 나오는 단어와 비교하였다.

기존 형태소 분석기에 비해 새로 구축된 형태소 분석기로 토큰화 하였을 때 공정가치, 법인세비용, 반기재무제표, 가중평균연간법인세율 등의 회계용어를 잘 인식하는 것을 확인하였다. 기존 형태소분석기를 활용해 “공정가치”를 토큰화 한다면 “공정”과 “가치”로 나누어 인식한다. 이는 “공정가치”라는 단어가 가지고 있는 정보를 없애기 때문에 정보전달이 주된 관심사인 재무제표의 목표를 해칠 뿐 아니라 향후 텍스트 마이닝 기법을 사용할 때에도 부적절하다.

5.2 시사점 및 연구의 한계

본 연구의 시사점은 다음과 같다. 본 연구에서는 재정정보를 활용하여 기존 형태소분석기를 회계용어 분석에 용이하게 튜닝 하였다. 주석정보의 중요성이 증가하고 기업들의 주석공시 내용의 유용성이 점차 증가할 것으로 기대되는 상황에서 새롭게 구축된 형태소 분석기는 재무제표의 텍스트 같은 회계용어가 포함된 텍스트를 분석할 때 기존 형태소 분석기보다 목적적합한 용어가 토큰화 될 수 있게 도울 수 있다.

텍스트 데이터는 인사, 재무, 마케팅 등 경영학과 관련된 여러 분야에 이미 사용되고 있지만 재무제표의 텍스트 데이터를 연구하는 것은 많이 이뤄지지 않았다. 하지만 재무제표를 비롯한 주석공시의 중요성이 증가되고 주석의 유용성 제고에 관한 연구가 이뤄지고 있음으로 향후 주석 데이터를 분석하는 것에 대한 수요가 있을 것이다. 본 연구에서 제안한 재무제표 분석을 위한 사전 구축이 향후 회계용어가 포함된 텍스트 데이터를 이용한 텍스트 마이닝 연구에 기여할 수 있을 것이고 유용한 정보를 외부이용자에게 전달하려는 재무제표의 본질을 살릴 수 있을 것이다. 새로 구축된 사전은 TF-IDF, Word2Vec, Doc2Vec, 토픽모델링, 감성분석 등의 텍스트 마이닝 기법들을 활용하기 위한 전단계로서 데이터를 전처리하는 과정에 사용될 수 있을 것이다. 분야별 LDA를 활용하여 기업들을 분류하고 각 토픽들이 가지는 특성을 파악해 볼 수도 있을 것이다(양낙영 등, 2019).

본 연구는 재정정보와 관련된 회계 용어를 중점적으로 분석하여 형태소 분석기를 구축하였다. 이 형태소 분석기에 회계용어 정보 뿐 아니라 정책관련 정보를 추가한다면 SNS를 분석하여 재정정보를 활용한 재정정책에 대한 국민들의 의견을 분석할 수도 있을 것이며 정책관련 문서를 분석할 수 있고 토픽모델링 기법을 통해 현재 어떤 유형의 정책이 실행되고 있는지 문서를 분류해 볼 수도 있을 것이다. 정부 및 공공기관의 민원관련 텍스트를 분석해 볼 수도 있을 것이다(김현종 등, 2018). 동일한 방법론으로 회계용어가 사용되는 분야가 아닌 다른 분야에 적용한다면 해당 분야에 적합한 용

어들을 형태소분석기를 활용하여 추출할 수 있을 것이다. 커뮤니티 게시판에서 주로 사용되는 용어를 바탕으로 여러 가지 텍스트 마이닝 방법론을 시도할 수도 있을 것이다(조혁준 등, 2017). 또한 자연어처리(NLP)를 위한 기계학습 데이터셋 구축에 사용될 수 있을 것이다. 기계학습 데이터셋 구축은 목적에 따라 데이터 수집 및 전처리 방법이 달라진다(윤종욱, 2019). 따라서 회계용어를 활용하는 목적을 가진 기계학습 데이터셋 구축에 사용될 수 있다.

본 연구의 한계점과 발전방향은 다음과 같다. 첫번째, 동음이의어 문제를 해결하지 못했다. '양산'은 도시이름이기도 하지만 '양산하다'의 원형이기도 하다. 위 단어가 어떤 의미를 가지는지 정확히 알기 위해선 앞 뒤 문맥을 파악해야 한다. 양산은 데이터 전처리 과정에서 도시명으로 불용어처리 하였기 때문에 문제가 발생하지 않았지만 비슷한 동음이의어 문제가 발생할 수 있다.

두번째, 유사어 사전을 추가하면 더 좋은 결과를 기대할 수 있다. 계정과목을 제외한 회계용어는 회사마다 사용 방식이 조금씩 다를 수 있다. 명사 사이의 조사나 종결어미의 사용에 따라 같은 의미의 단어가 다르게 표현될 수 있다. 본 연구에서는 이에 대한 해결방안으로 모두 명사로 토큰화 한 다음 N-gram을 적용하였지만 모든 경우를 해결하지 못했다. 예를 들어 "가중평균을 적용한 연간법인세율" 과 '가중평균연간법인세율'은 같은 의미이지만 "가중평균을 적용한 연간법인세율"을 명사로 토큰화 한 후 N-gram을 적용한다면 "가중평균적용연간법인세율"이라는 결과가 나온다. 이를 해결하기 위해 비슷한 의미의 단어를 한 단어로 치환할 수 있는 정규화를 추가한다면 유사어와 관련된 문제를 해결할 수 있을 것이다.

참고문헌

강승식, "sms 영역에 대한 형태소 분석 사전의 구축," 언어정보, 9권, 2008, pp. 5-21.

권희준, 김선우, 임을규, "Multi N-Gram을 이용한 악성코드 분류 시스템," 보안공학연구논문지, 9권, 6호, 2012, pp. 531-542.

모예린, 서윤석, "주석 내용의 변동과 주식시장 : 주석 내용의 변동이 자기자본비용과 주식거래량 및 이익반응계수에 미치는 영향," 회계학연구, 44권, 4호, 2019, pp. 215-249.

박경진, 김기영, 송문섭, "k-Ifrs하에서 오류수정으로 인한 재무제표 재작성의 문제점 - 사례분석을 중심으로," 회계저널, 23권, 2호, 2014, pp. 345-368.

박기영, 이영준, 김수현, "텍스트 마이닝을 활용한 금융통화위원회 의사록 분석" The Korean Economic Review, 35권, 2호, 2019, pp. 471-511.

송은지, "소셜 미디어 상 고객피드백을 위한 감성분석," 한국정보통신학회논문지, 19권, 4호, 2015, pp. 780-786.

심광섭, "품사 태깅 말뭉치에서 추출한 n-gram을 이용한 음절 단위의 한국어 형태소 분석," 정보과학회논문지: 소프트웨어 및 응용, 40(12), 2013, pp. 869-876.

김정호, 김명규, 차명훈, 인주호, 채수환, "한국어 특성을 고려한 감성 분류," 감성과

학, 13(3), 2010, pp. 449-458.

이정민, 전은자, 채정민, “텍스트 마이닝을 기반으로 한 무용학 자료의 빅데이터 분석,” 무용역사기록학, 42권, 2016, pp. 191-212.

이현영, 이종석, 강병도, 양승원, “효율적인 한국어 파싱을 위한 최장일치 기반의 형태소 분석기 기능 확장,” 한국디지털콘텐츠학회 논문지, 17권, 3호, 2016, pp. 203-210.

장진영, “회계 전문용어의 언어학적 분석,” 언어과학연구, 60권, 2012, pp. 191-212.

황선필, 윤재원, 김경호, “K-IFRS 도입 전후의 주석공시사례분석과 전문가 평가,” 상업교육연구, 31권, 2호, 2017, pp. 179-205.

양낙영, 김성근, 강주영, “텍스트 마이닝 방법론과 메신저 UI 를 활용한 융합연구 촉진을 위한 연구자 및 연구 분야 추천 시스템의 제안,” 정보시스템연구, 27(4), 2018, 71-96.

조혁준, 김성근, 강주영, “도플갱어 브랜드 이미지 효과에 대한 실증적 분석: 인터넷 커뮤니티를 중심으로,” 정보시스템연구, 26(1), 2016, pp. 21-51.

한국회계기준원(KAI), 재무보고를 위한 개념체계, 2016

이원희, “4차 산업혁명과 5g 시대의 재정정보 관리,” 재정포럼, 274, 2019, pp. 2.

윤종욱, “기계학습 데이터세트 구축 공정 표준화에 관한 파일럿 연구,” 인터넷전자상거래연구, 19권, 제5호, 2019, pp. 199-217.

김현종, 이태헌, 유승의, 김나랑, “민원 분석을 위한 텍스트 마이닝 기법 연구: 계층적 연관성 분석,” 한국산업정보학회논문지, 23권, 제3호, 2018, pp. 13-24.

Beaver, William H, “Financial Ratios as Predictors of Failure,” *Journal of Accounting Research*, 1966, pp. 71-111.

Tripathy, A., Agrawal, A., Rath, S. K., “Classification of sentiment reviews using n-gram machine learning approach,” *Expert Systems with Applications*, 57, 2016, pp. 117-126.

DECHOW, PATRICIA M., WEILI GE, CHAD R. LARSON, and RICHARD G. SLOAN, “Predicting Material Accounting Misstatements,” *Contemporary Accounting Research*, Vol.28, No.1, 2011, pp. 17-82.

Dutta, Shantanu, Ila Dutta, and Bijan Raahemi, “Detecting Financial Restatements using Data Mining Techniques,” *Expert Systems with Applications*, Vol.90, ,2017, 374-393.

Kamaruddin, Siti Sakira, Azuraliza Abu Bakar, Abdul Razak Hamdan, Fauzias Mat Nor, Mohd Zakree Ahmad Nazri, Zulaiha Ali Othman, and Ghassan Saleh Hussein, “A Text Mining System for Deviation Detection in Financial Documents,” *Intelligent Data Analysis*, Vol.19, No.s1, 2015, pp. S19-S44.

Ravisankar, P., Ravi, V., Raghava Rao G., and

Bose, I., “Detection of Financial Statement Fraud and Feature Selection using Da ta Mining Techniques,” *Decision Support Systems*, Vol.50, No.2, 2011, pp. 491-500.

Pérez, Carmen Caba, Antonio M. López Hernández, and Manuel Pedro Rodríguez Bolívar, “Citizens’ Access to on-Line Governmental Financial Information: Practices in the European Union countries,” *Government Information Quarterly*, Vol.22, No.2, 2005, pp. 258-276.

Perkins, Jacob. Python 3 text processing with NLTK 3 cookbook. *Packt Publishing Ltd*, 2014.

Post, Matt, and Shane Bergsma. “Explicit and implicit syntactic features for text classification.” Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2013.

http://www.openfiscaldata.go.kr/portal/baeoom/baeoom01.do(2019.09.30.)

https://www.samili.com/acc/kijun/Kifrs_tot_list.asp?op=1&op2=1(2019.09.30.)

https://wikidocs.net/21692(2019.11.30.)

https://web.stanford.edu/~jurafsky/slp3/3.pdf(2019.11.30.)

정 건 용 (Jung, Geon-Yong)

아주대학교 e-비즈니스학과에서 재학 중에 있다. 주요 관심분야는 텍스트 마이닝, 머신러닝, 딥러닝 등 이다.

윤 승 식 (Yoon, Seung-Sik)

아주대학교 e-비즈니스학과에서 학사학위를 취득하였다. 현재 스노우에 근무하고 있으며 주요 관심분야는 텍스트 마이닝, 머신러닝, 딥러닝 등 이다.

강 주 영 (Kang, Ju-Young)

현재 아주대학교 경영대학 e-비즈니스학과 교수로 재직 중이며, 포항공과대학교 컴퓨터공학과에서 학사, 서울대학교 컴퓨터공학과에서 석사, 한국과학기술원 경영공학전공에서 공학박사학위를 취득하였다. 주요 관심분야는 빅데이터, 텍스트 마이닝, 시맨틱 웹, 지능형 정보시스템 등이다.

<Abstract>

Development of Text Mining-Based Accounting Terminology Analyzer for Financial Information Utilization

Jung, Geon-Yong · Yoon, Seung-Sik · Kang, Ju-Young

Purpose

Social interest in financial statement notes has recently increased. However, contrary to the keen interest in financial statement notes, there is no morphological analyzer for accounting terms, which is why researchers are having considerable difficulty in carrying out research. In this study, we build a morphological analyzer for accounting related text mining techniques. This morphological analyzer can handle accounting terms like financial statements and we expect it to serve as a springboard for growth in the text mining research field.

Design/methodology/approach

In this study, we build customized korean morphological analyzer to extract proper accounting terms. First, we collect Company's Financial Statement notes, financial information data published by KPFIS(Korea Public Finance Information Service), K-IFRS accounting terms data. Second, we cleaning and tokeninzing and removing stopwords. Third, we customize morphological analyzer using n-gram methodology.

Findings

Existing morphological analyzer cannot extract accounting terms because it split accounting terms to many nouns. In this study, the new customized morphological analyzer can detect more appropriate accounting terms comparing to the existing morphological analyzer. We found that accounting words that were not detected by existing morphological analyzers were detected in new customized morphological analyzers.

Keyword: text mining, text tokenizing, N-gram, morphological analyzer, accounting terminology analyzer, financial statements, notes

* 이 논문은 2019년 11월 18일 접수, 2019년 11월 29일 1차 심사, 2019년 11월 29일 게재 확정되었습니다.