SNS 데이터와 Word2Vec을 이용한 게임 콘텐츠 평가

Evaluation of game contents using SNS data and Word2Vec

김익중[1], 김정윤[2]

Ik-Joong kim[1] , Jung-Yoon kim[2]

요약

SNS는 각종 콘텐츠들에 대한 사용자들의 평가가 생성되는 곳으로 평가를 추론하는 데이터로 유용한 곳이다. 그러나 SNS 데이터는 비속어를 비롯해 채팅 용 문장들이 많으며 문법이 정형적이지 않아 기계적으로 의미를 해석하기가 어려움이 있다. 본 논문에서는 특정 게임 사용자들을 대상으로 SNS 데이터를 6개월간 취합하여 Word2Vec을 이용하여 게임 콘텐츠가 사용자들에 가진 의미를 분석하였다. SNS 데이터는 CBOW과 Skip-gram을 비롯하여 문장 간 거리를 비교하였으며 SNS 데이터에 대한 합리적인 기계 학습 모델을 제안한다.

핵심어: 자연어처리, Word2Vec, CBOW, 비지도 학습, SNS

Abstract

SNS is useful place for infer evaluation based on user generated feedback regarding various contents.However, SNS data contains many profanity and internet slangs, and wrong grammar. So it makes difficult to understanding and analyze meaning by the computer. In this paper, we collected SNS data for specific game users for 6 months. Then, Word2Vec was used to analyze the meaning of game contents to users. SNS data compares the distance between sentences including CBOW and Skip-gram, and suggests a reasonable machine learning model for SNS data.

Keyword: NLP, Word2Vec, CBOW, Unsupervised Learning, SNS

1) Graduate School of Game, Gachon University, 1342 Seongnam Daero, Sujeong-Gu, Seongnam-Si, Gyeonggi-Do, South Korea, 461-701 e-mail: gamedev@gc.gachon.ac.kr (First author)
2) Graduate School of Game, Gachon University, 1342 Seongnam Daero, Sujeong-Gu, Seongnam-Si, Gyeonggi-Do, South Korea, 461-701 e-mail: kjyoon79@gmail.com (Corresponding author)
Received(April 22. 2018), Review (May 16. 2018), Accepted(June 30. 2018)

1. 서론

SNS는 각종 콘텐츠나 서비스에 대한 즉흥적인 감정 표출이 많은 곳이다. 특히 트위터는 다른 SNS보다 문장이 짧으며 다른 사용자들의 즉각적인 피드백을 이끌어내는 특징이 있다[1]. 그러나 SNS의 데이터는 비속어와 욕설이 함께 사용되는 경향이 높으며 문법적으로 틀린 문장들이 많기에[2] 단순 사전이나 문장 분류로는 의미 파악이 어렵다. 또한 비속어와 욕설이 많다고 해서 반드시 해당 콘텐츠에 부정적인 평가인 것도 아니다. 비속어와 함께 콘텐츠를 긍정적인 평가를 내리는 경우도 있으며[3], 해당 브랜드의 이해나 충성도가 높을수록 SNS에서의 의견 표출 역시 높아지는 경향을 보여준다[4]. 무엇보다 콘텐츠를 생산한 측에서 제시한 명칭과 달리, 사용자 측에서 자생한 축약어와 별명 등이 빈번하게 사용된다. 따라서 SNS의 데이터를 분석하기 위해서는 기존의 문장 분류 체계 및 콘텐츠 생산자가 제공하는 특정 단어만으로는 불가능하다.

예컨데 게임 "함대콜렉션 칸코레"는 "깡게임", "깡겜"등의 된소리로 사용되며 "소녀전선"은 "소전"으로, "이벤트"는 "이벤"으로 축약된다. "아이돌 마스터 신데렐라 걸즈"는 "신데마스" 혹은 "데레마스" 등으로 변형되어 사용된다. 문제는 이런 변형된 자료들과 신조어들은 기존 형태소 분류 사전에 등록되지 못했기에 품사 분류 과정에서 사라지는 경우가 많다. "칸코레"는 "칸|코레"로 분류되며 "신데마스"는 "신데|마스"로 분류되어 핵심 단어가 기계학습에서 누락되는 것이다.

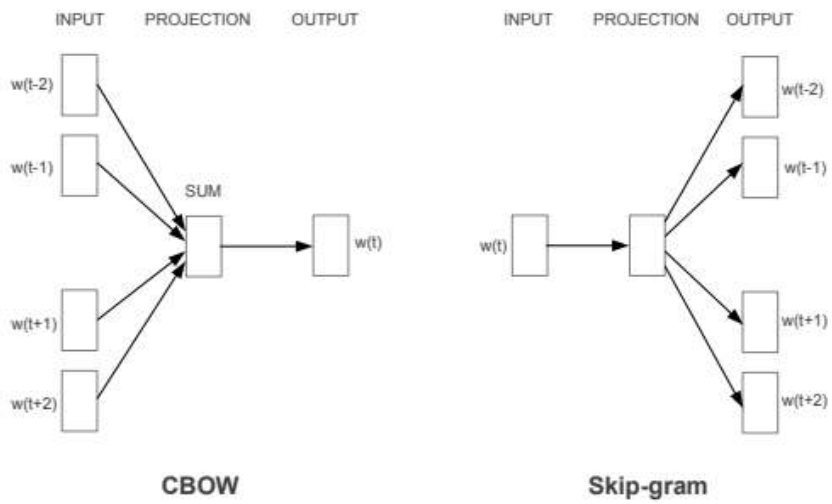
이에 SNS 데이터는 두단계로 학습시켜볼 필요가 있다. 우선은 품사를 분리하지 않은 형태에서 학습을 통해 사용자들이 실제 사용하는 단어를 찾아낸 후, 해당 단어를 다시 형태소 사전에 등록시켜 분리된 단어를 중심으로 의미를 추론하여야 한다. 의미 추론에는 Word2Vec[5]을 사용하였으며 알고리즘은 CBOW와 Skip-gram을 함께 사용하였다. 또한 SNS 데이터는 문장구조가 짧다는 특성에 기반 하여 인접 거리를 각각 다르게 설정하여 학습시켜 보았다.

본 논문의 구성은 다음과 같다. 2장에서는 이론적 배경과 선행 연구들을 살펴보며 3장에서는 실험을 위한 제안과 과정을 소개하며 4장에서는 실험 결과를 도출한다. 5장에서는 실험 결과가 가지는 의미와 앞으로의 연구 방향에 대해 제시한다.

2. 이론적 배경 및 선행 연구

2.1 Word2Vec

인접한 단어들은 통계적으로 높은 관계성이 가진다는 언어학의 이론[6]을 바탕으로 신경망을 접목시킨 NNLM[7, 8]과 RNNLM[9]가 2000년대에 등장하였다. 이 알고리즘들은 단어 간의 거리를 벡터화 하여 단어끼리의 거리를 통해 유사어를 추론하지만 실용적인 학습을 위해 신경망 학습 시간을 단축시킬 방법이 필요해졌다. Word2Vec은 2013년에 Tomas Mikolov가 발표한 알고리즘으로 기존 알고리즘보다 개선된 속력을 가진 CBOW(Continuous Bag of Words Model)와 Skip-gram 알고리즘을 동시에 지원한다.



[그림 1] CBOW와 Skip-gram]

[Fiq. 1] CBOW and Skip-gram]

CBOW는 문장 사이에서 중간에 들어갈 단어를 추론할 수 있는 알고리즘이다. {"The", "cat", 'over", "the', "puddle"}에서 "over"와 "the"사이에 "jumped"를 추론할 수 있어야 한다. 앞선 문장을

$$(x^{(c-m)}, \ldots, x^{(c-1)}, x^{(c+1)}, \ldots, x^{(c+m)} \in \mathrm{R}^{[v]})$$

로 표현할 때 "jumped"는 로 학습시키며 문장의 크기는 m으로 한다. 이 경우 "jumped"의 벡터는 다음과 같다.

$$(v_{c-m} = Vx^{(c-m)}, v_{c-m+1} = Vx^{(c-m+1)}, \dots, v_{c+m} = Vx^{(c+m)})$$

여기에 벡터의 평균값이 를 얻는다.

$$\hat{v} = \frac{v_{c-m} + v_{c-m+1} + \dots + v_{c+m}}{2m}$$

벡터 z의 점수는 $z = U\check{v}$로 표현하며 z에서 확률 $\check{y}$를 구한다.

$$y = softmax(z)$$

이를 목적함수로 정리하면 다음과 같다.

$$\begin{aligned} minimize\ j &= -\log P(w_c | w_{c-m}, \dots, w_{c-1}, w_{c+1}, \dots, w_{c+m}) \\ &= -\log P(u_c | \hat{v}) \\ &= -\log P \frac{\exp(u_c^T \hat{v})}{\sum_{j=1}^{|v|} \exp(u_j^T \hat{v})} \\ &= -u_c^T \hat{v} + \sum_{j=1}^{|v|} \exp(u_j^T \hat{v}) \end{aligned}$$

중심단어가 x^c가 될 확률은 Word2Vec이 만든 두개의 벡터 u_c와 y_c사이에서 거리를 감소해가며 계산할 수 있다.

Skip-gram은 CBOW와 반대의 성격을 가진다. "jumped"가 주어지면 학습한 주변 단어인 {"The", "cat", 'over", "the', "puddle"}을 찾아준다. 목적함수는 다음과 같으며 순회를 하며 입력된 단어의 범위를 좁혀간다.

$$\begin{aligned} Minimize\ J &= -\log P(w_{(c-m)}, \dots, w_{(c-m)}, w_{(c+m)}, \dots, w_{c+m} | w_c) \\ &= -\log \prod_{j=0, j\neq m}^{2m} P(w_{c-m+j} | w_c) \\ &= -\log \prod_{j=0, j\neq m}^{2m} P(w_{c-m+j} | v_c) \\ &= -\log \prod_{j=0, j\neq m}^{2m} \frac{\exp(u_k^T v_{c-m+j} v_c)}{\sum_{k=1}^{|V|} \exp(u_k^T v_c)} \\ &= -\sum_{j=1, j\neq m}^{2m} u_k^T v_{c-m+j^{v_c}} + 2m \log \sum_{k=1}^{|V|} \exp(u_k^T v_c) \end{aligned}$$

2.2 MeCab-ko

MeCab[10]은 2006년에 개발되어 2013년에 안정화된 일본에서 발표된 형태소 분리기이다. MeCab-ko는 일본어와 한국어가 비슷하다는 점에 착안하여 국내의 은전한닢 프로젝트를 통해 한국어 사전이 만들어진 분석기이다. 한국어 형태소 분석기는 Twitter, KKMA, Komoran, Hannanum 등이 있으나 본 실험에서는 Mecab-ko를 선택하였다. C 언어로 제작되어 플랫폼 특성을 나타내지 않는다는 점과 수행 시간이 다른 분석기들 빠르다는 점이 선택 이유였다. Mecab-ko은 21세기 세종 계획 사전 및 말뭉치를 이용하여 정확도가 높으며 사용자가 사전을 추가할 수 있어 게임 타이틀이나 은어와 같은 단어들을 학습시킬 수 있었다. [표1]은 Mecab-ko와 기본 사전을 통한 품사 분리 형태이다.

[표1 MeCab-kr의 결과 사례]

[Table 1 Result of MeCab-kr]

입력	오늘은 날씨가 무척 좋다.	
출력	EOS	
		SY,*,*,*,*,*,*,*
	오늘	NNG,*,T,오늘,*,*,*,*
	은	JX,*,T,은,*,*,*,*
	날씨	NNG,*,F,날씨,*,*,*,*
	가	JKS,*,F,가,*,*,*,*
	무척	MAG,성분부사/정도부사,T,무척,*,*,*,*
	좋	VA,*,T,좋,*,*,*,*
	다	EF,*,F,다,*,*,*,*
	.	SF,*,*,*,*,*,*,*
	EOS	

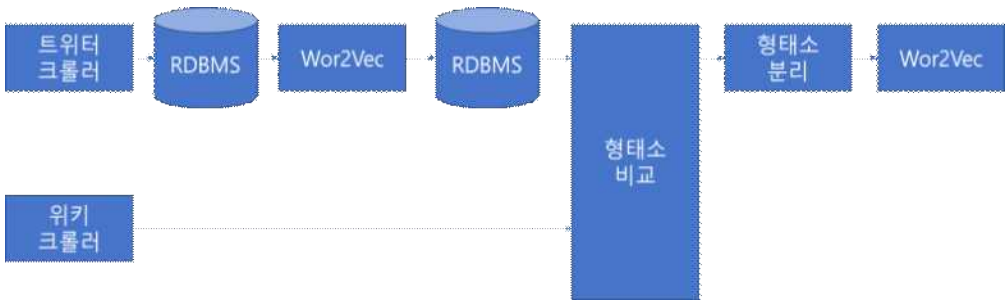
한국어의 특성상 학습을 시키기 이전에 조사를 분리해야 하지만 SNS 데이터라는 특징상 다양한 명사 이외에 다양한 품사를 조사할 필요가 있었다. 오피니언 분석에서 명사만 추출하면 단어가 지니는 감정과 평가에 대해 누락될 수 있기 때문이었다. [11] 따라서 이 실험에서는 나무위키

와 SNS 데이터의 품사 개수를 비교하여 평가가 될 데이터의 품사 분류에 대한 기준을 세웠다.

3. 실험 제안 및 과정

3.1 실험 구축

실험 데이터를 모으기 위해 우선 게임 평가에 대한 트윗이 많은 이용자들을 선별해 그들의 트윗을 크롤링하였다. 크롤링된 데이터는 RDBMS로 저장되며 Full Text Search를 사용하여 수시로 조사할 수 있었기에 크롤링 중에도 정상적인 데이터들이 수집되고 있는지를 확인할 수 있었다. 데이터의 빠른 처리를 위해 크롤러와 데이터 가공 프로그램은 C#으로 제작하였으며 형태소 분석기와 Word2Vec 관련은 C/C++로 제작하였다. 학습을 시키기 위한 데이터 크기는 GB 단위였기 때문에 64비트로 빌드하였다.



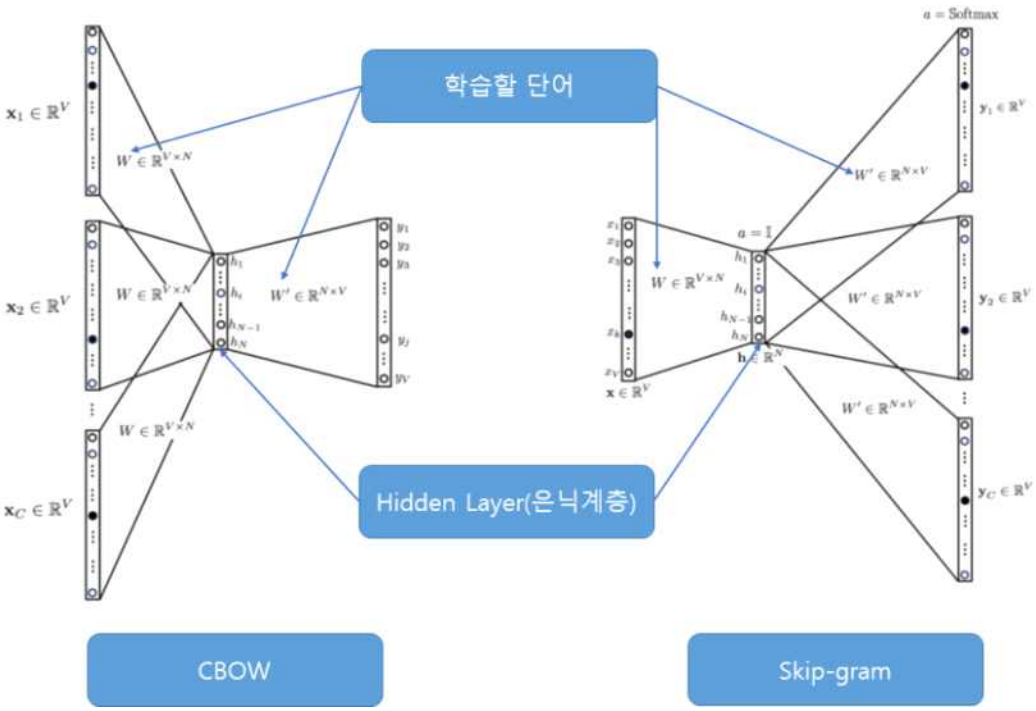
[그림 2] 분석 시스템 구조도

[Fig. 2] System Structure for Analysis

크롤링 기간은 2018년 1월부터 6월까지였으며 매달 품사를 분류하지 않는 상태에서 비지도 학습을 통해 특정 게임에 대한 근접 단어를 파악하였다. 데이터 수집 결과로 나온 근접 단어를 형태소 사전에 추가한 후, 다시 품사를 분류해 비지도 학습의 대상으로 하여 콘텐츠의 평가를 시행하였다.

Word2Vec은 CBOW와 Skip-gram이라는 두 가지 학습 알고리즘을 지원한다. CBOW는 주어진 단어에 대해 앞 뒤로 C/2씩 총 C개의 단어를 만들어 네트워크를 구성한다. 이 결과 "칸코레 이

벤트를 하는데 ____가 없어서 제공권이 모자르다."라는 문장에서 ____가 주어지지 않아도 문장을 유추하는데 도움을 준다. Skip-gram은 CBOW와 반대로 주어진 단어를 기준으로 주위 단어를 예측하는 기법이다. 가까이 있는 단어가 높은 관계성을 지니는 단어일 것이라 가정하고 멀리 떨어져 있는 단어일수록 낮은 확률로 선택한다. 이는 한 단어가 가지는 두 가지 의미를 학습할 수 있다. 즉, "애플"이라면 과일을 의미하는지, 회사 이름을 의미하는지를 파악하는데 유의미한 알고리즘이다.



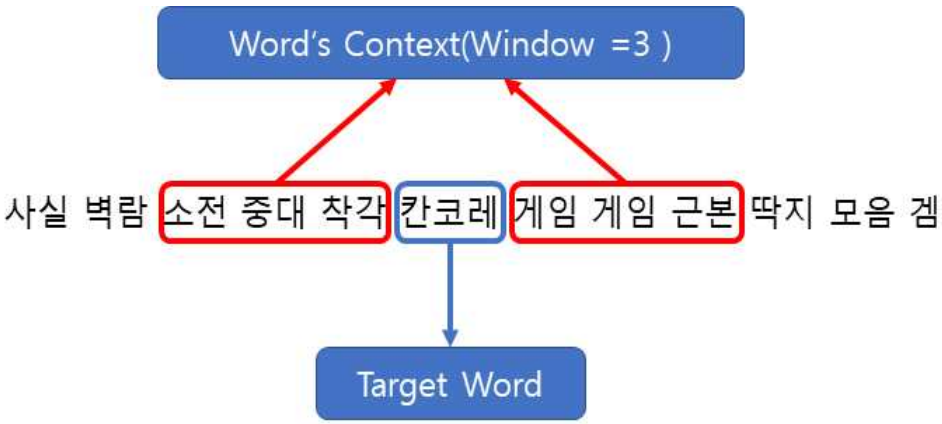
[그림 3] 실험에 사용된 CBOW와 Skip-gram 비교

[Fig. 3] Comparison of CBOW and Skip-gram used in the experiment

알고리즘의 특징으로 Skip-gram은 CBOW보다 학습량이 많기에 더 정확한 의미를 지닐 수 있기에 대부분의 경우 Skip-gram을 선호하는 실정이지만 Skip-gram은 그만큼 학습 데이터가 많아야 효율이 높아진다는 점을 간과해서는 안된다. 이 실험에서는 단어의 유사성과 사용자의 평가가 함께 연구되어야 하기에 CBOW와 Skip-gram을 각각 사용하여 결과를 살펴보았다.

또한 주위 단어 범위인 윈도우(Context Window)를 5와 12로 설정하여 유의미한 결과를 찾는

데 주력하였다. 다른 데이터들에 비해 단문이 많은 트위터의 특성과 한국어의 특성을 고려한 실험이었다. [그림 4]는 학습 시 주위 단어를 파악하는 윈도우 개수에 따른 사례를 보여준다.



[그림 4] 윈도우 사이즈 3일 경우의 Context 학습 범위

[Fig. 4] Context learning range for window size 3

3.2 품사 선택 기준

2018년 4월 16일 통계로 나무위키는 575661개의 문서를 가지고 있으며 게임 콘텐츠에 대한 분량이 위키피디아에 비해 많은 편이다. 콘텐츠나 게임에 대한 설명과 함께 사용자들의 평가도 언급되어 있기에 품사 분류를 위한 문장 구조 파악 설계에 참고자료로 사용하였다.

[표2] 나무위키와 트위터 데이터의 품사 순위 비교

[Table 2] Frequency comparison of Namu-Wiki and Twitter

나무위키		SNS(트위터)	
일반 명사	39%	일반 명사	32%
고유 명사	5%	연결 어미	8%
부사격 조사	5%	동사	7%
연결 어미	5%	마침표, 물음표, 느낌표	5%
구분자	5%	일반 부사	5%
보조사	4%	고유 명사	5%

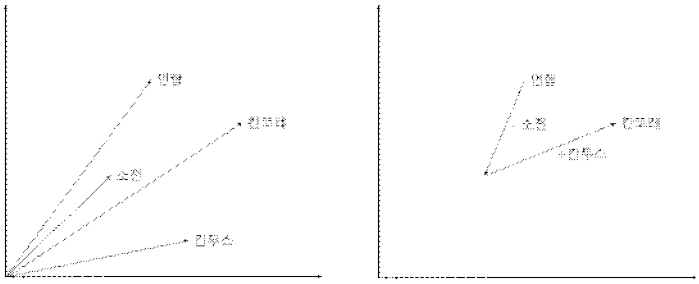
동사	4%	보조사	4%
마침표, 물음표, 느낌표	3%	부사격 조사	4%
목적격 조사	3%	관형형 전성 어미	3%
주격 조사	3%	주격 조사	3%
관형형 전성 어미	3%	명사 파생 접미사	3%
일반 부사	3%	의존 명사	3%
의존 명사	2%	종결 어미	2%
명사 파생 접미사	2%	감탄사	2%
관형격 조사	2%	목적격 조사	2%
종결 어미	2%	형용사	2%
단위를 나타내는 명사	2%	관형사	2%
동사 파생 접미사	2%	동사+연결 어미	2%
형용사	2%	보조 용언	2%
관형사	1%	대명사	2%
보조 용언	1%	동사+관형형 전성 어미	1%
동사+관형형 전성 어미	1%	동사 파생 접미사	1%
동사+연결 어미	1%	긍정 지정사	1%
접속 조사	1%	관형격 조사	1%
합계	100%	합계	100%

[표2]는 나무위키와 SNS에서 사용되는 품사 중 1% 이상인 것들을 비교한 것이다. 나무위키는 URL에 /raw/를 추가하여 HTML이 아닌 TEXT 형태로 크롤링 하였다. raw 형식의 문서에서 링크 구분자인 "|"등의 문장 부호를 제외시켰으며 SNS 데이터 역시 외국인 사용자가 입력한 외래어를 비롯하여 원하는 문장 정보와 관련성이 적은 품사 등을 제외시켰다. 두 케이스 모두 일반 명사가 30% 이상을 차지하지만 2순위부터 확연한 상호 차이를 보인다. 특히 SNS에는 연결 어미, 동사와 감정을 나타내는 물음표와 느낌표가 높은 순위를 차지하는 것은 행위에 대한 설명한 감정이 함께 나타나는 상당히 유의미한 지표라 할 수 있을 것이다.

3.2 Word Analogy

다수의 차원으로 벡터 거리가 규명된 단어들은 벡터 연산(Vector Representation)을 통해 연관된 단어를 추론해 볼 수 있다. 학습이 성공하였다면 "King – Man + Woman = ?"이란 질문에

"Queen"이라고 대답할 수 있어야 한다. 이 실험에서는 게임 콘텐츠에 대한 실험이므로 "A 게임의 유닛명 – A 게임의 타이틀명 + B 게임의 유닛명"을 질문하였을 때 "B 게임 타이틀명"을 대답하면 성공적인 학습이었다고 가정한다. [그림 3]은 "인형 – 소전(소녀전선) + 칸무스" 일 경우 "칸코레"를 찾아갈 수 있는 벡터를 표현한 것이다. 또한 사용자들은 게임을 하며 자신의 감정을 나타내는 경향이 있으므로 감정을 나타나는 단어를 입력하며 게임 콘텐츠와 가까운 단어를 추론할 수 있도록 하였다.



[그림 5] 단어와 벡터 연산

[Fig. 5] Vector Representation

4. 실험 및 결과

4.1 기간별 CBOW 학습 결과

게임을 하는 SNS 사용자는 게임에 대한 감상을 SNS에 적는다는 가설 아래 핵심 단어를 추출하기 위해 기간별로 학습결과를 확인하였다. 감정이 담긴 비속어나 감탄사 등도 함께 조사하기 위해 CBOW 알고리즘으로 학습시켰으며 윈도우 사이즈는 5로 설정하였다.

[표 3] 2018년 2월 데이터로 분석한 근접 단어

[Table 3]Similarity Distance in February 2018

단어	근접 단어
게임	모바일, 개발자, 최근, 비디오, 방볶음, 짱, 한국에, 작은, 콘솔, 플레이, 화가, 관련해서, 오브, 앱, 다리, 설치, 엠바고, 퍄
이벤트	제목이, 마지막으로, 왔다, 데레스테, 한섭
소전	젤다, 콜라보, 걸판, 흐음, 왜케, 귀여워
소녀전선	업데이트, 한국어, 공유, 군생활, 기사
벽람	벽람, 넵튠, 성능이, 배터리, 병신, ㅜㅜ, 상태가, 하스, 경쟁
벽람항로	
깡겜	
칸코레	반응이, 맨날, 오랜만에, 쓰레기, 겜, 함, 들어가면, 아닌가, 살까, 좋지, 없네
재밌다	사이가, 몬헌월드, 판정, 전형적인, 부족해서
재밌는	
시발	
씨발	존나, 씨발, ㅋㅋㅋㅋㅋㅋㅋ, ㅋㅋㅋㅋㅋㅋ, 와, ㅋㅋㅋㅋ, ㅋㅋㅋ, ㅋㅋㅋㅋㅋㅋㅋㅋ, ㅅㅂ, ㅋㅋㅋㅋㅋㅋ, ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ, 시바, 미친, ㅋㅋㅋㅋㅋㅋㅋㅋ, ㅋㅋㅋㅋㅋㅋㅋㅋ, ㅁㅊ, ㅋㅋㅋㅋㅋ
소울워커	

[표 4] 2018년 3월 데이터로 분석한 근접 단어

[Table 4] Similarity Distance in March 2018

단어	근접 단어
게임	콘솔, 게임을, 개발자,게임이, 회사가, 애니메이션, 게임은, 유저, 모바일, 서비스, 기억에, 시스템, 플레이, 기능, 오너가, 시리즈, 글로벌, 신규, 하스스톤
이벤트	이벤트, 콜라보, 시즌, 진행, 이번, 특이점, 설에, 이벤트가, 보상으로, 멀린, 데레스테, 소전, 이번에, ㅋㅋㅋ, 그랑블루, 서번트, 라멘, 5위, 유저, 예장
소전	소전, 칸코레 이벤, 포인트, 만들까, 이벤트가, 요약, 갓겜, 아.., 재밌네, 데레스테, 랭킹전, 신, 히든, 이벤트, 스토리, 후.., 재밌다, 방도리, ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
소녀전선	콜라보, 만화. 걸판겜. 미쿠, feat, 우에다, 1주년, MAP, 업데이트, 머야, 주간, 굿즈, 시리즈, 애플, 코어
벽람	벽람, 칸코레, 그랑블루, 할거, 재밌다, 이벤, 우와, 힘들다, 넵튠, 망했네, 한섭
벽람항로	스크린샷, 다행이야, 보드게임, 라이트, 신규
깡겜	나와?, 왤캐, 개웃기네, ㅊㅋㅊㅋ, 108호
칸코레	벽람, 소전, 이벤트가, 칸코레는, 그랑블루, 특이점, 섭종, 근황, 이벤, 요약, 몰랐네, 1200만, 겜, 일섭, 중섭, 발표, 랭킹전, 으이구, 콜라보는
재밌다	힘들다, ㅜㅜ, 조아, 이쁜데, 이쁘다, 귀엽다, 재밌네, 겁나, 보이스, 짜증난다, 좋아, 귀여움, 먹고싶다, 힘드네, ㅁㅊㄷ, 으아, 한번만, 웃김
재밌는	어려운, 것이라는, 파는, 여유가, 중요한, 의외로, 쉽게, 아니냐, 인기가, 정도로, 느껴지는
시발	씨발, 씨벌, 시벌, 웃기네, ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ, 씹, 시바, 새끼들아, 이지랄, 쉬벌, 맞아, 염병, ㅋㅋㅋㅋㅋㅋㅋ, 아나, 나이트메어, 미친
씨발	씨발, 씨벌, 애미, 시발, 병신, 처, 새끼들아, 이러니까, 씹새끼들, 틀딱, 좆같은, 새끼, 병신같은, 버러지, 씨팔, 새끼들이, 씨발년들아, 맞아, 죠센징, 염병
소울워커	

[표 5] 2018년 6월 데이터로 분석한 근접 단어

[Table 5] Similarity Distance in June 2018

단어	근접 단어
게임	게임을, 모바일, 개발자, 게임이, 게임으로, 유저, 플레이, 모바마스, 룰, 게임, 게임은, 콘솔, 스트리밍, 시스템, 콘텐츠가, 클래스, 개발, 게임에, 워런티
이벤트	페그오, 한정, 콜라보, 그랑블루, 스토리, 보상, 특이점, 파밍, 픽업, 난이도, 이벤, 다운로드, 이벤트도, 벽람, 보상으로
소전	벽람, 콜라보, 칸코레, 소울워커, 방도리, 오버워치, 일러, K7, 반도리, 서버, 고인물, 페그오, 유저들, 패키지, 요약, 4주년, 블소, 히오스, 일러레
소녀전선	1주년, 오버워치, 스팀, 케이온, 유저, 보이스, 정박아, 샤니마스, 로그인, 1화, 콜라보, 던전, 뱅드림, 포스터, 10주년, 디비전, 스토리, 유튜브에서, SSR
벽람	소전, 칸코레, 소울워커, 콜라보, 듀랑고, 한섭, 이벤, 샤니마스, 나왔나, 갓겜, 겜, 방도리, 일러, 난이도, 페그오, 퐛, 마비, 그랑블루, 전함소녀
벽람항로	패션, 발매일이, 콜라보, 추가!, 토오루, 1주년, 포켓몬스터, 이치방쿠지, 한국어판, 스킨, 일러스트, 3회차, 2기, 2020년, 수영복, 이벤트와, 1권, 기념
깡겜	신성, 아아니, 왜이래, 에반데, 자랑이라고, 너무하네, 실화?, 꿀잼이네, 막판에, 일어났다, 아이엠스타, 과금하면, ㅂㅂㅂㄱ, 힘드네, ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ, 소워, 떴다
칸코레	소전, 벽람, 오버워치, 소울워커, 듀랑고, 칸코레는, 일러, 넵튠, 뉴비, 콜라보, 블리자드, 4주년, 클로저스, 갓겜, 스토리가, 서버, 페그오, 매출, 스팀
재밌다	귀엽다, 재밌어, 예쁘다, 웃겨서, 귀여운데, 이뻐, 귀엽네, 이쁘네, 이뻐서, 춥다, 커엽, 심하다, 안구사, 좇같음, 이쁘다, 좋았는데, 웃겨, 조아
재밌는	어려운, 무서운, 무척, 야한, 행복한, 편한, 바꾸는, 이세계에, 재미있는, 틀린, 진지한, 느끼는, 그런, 배운, 안다는, 귀엽다고, 읽은, 많다는, 여러모로
시발	씨발, 시팔, ㅅㅂ, 했지, 시벌, 퇴물, 처, 시바, 쉬벌, 저기요, 저사람, ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ, 씨팔, 염병, 하네, 씹, 진짜?
씨발	시발, 씨벌, 씨팔, 처, 좇같은, 좇같다, 병신들아, 애미, 즈그, 개새끼야, 병신같은, 그 새끼, 좇같네, ㅉㅉㅉ, 개새끼들아, 새끼들아, 지랄을, 염병
소울워커	소울워커, 소전, 벽람, 그랑블루, 이벤, 방도리, 칸코레, 콜라보, 아제로스, 마비, 한섭, 4주년, 장담하는데, 요약, 중섭, 본편, 안함, 클로저스, 보상으로, 페그오

[표 4]에서 보이는 것처럼 3달간 수집된 데이터에서 비슷한 성격의 게임이 SNS 사용자들에게 언급되고 있었다. 실제 게임 플레이에 무관하게 해당 게임을 알고 있거나 관심이 있다고 봐야할 것이다. 또한 게임의 이벤트와 맞물려 이벤트와 보상, 새로운 캐릭터의 일러스트에 대해 논의함도 확인되었다. 단순 인접 단어들만으로도 실제 게임 운영에 도움을 줄 지표들을 찾아낼 수 있음을 확인할 수 있었다. 특히 [표 5]의 "소울워커"처럼 특정 시기에 사회적으로 화제가 된 게임은 다른 게임을 언급하던 사용자들에게서 높은 연관성을 가지는 것이 확인되었다.

4.2 품사 분리와 벡터 연산을 통한 평가 분석

2018년 6월 데이터로 SNS 사용자의 문장에서 원형 손실을 최대한 막고 단계에서 형태소 분석을 수행하였다. 이 때 corpus에 저장한 품사는 감탄사 숫자, 형용사+연결어미, 단위 명사, 형용사+관형형 전성 어미, 일반 부사, 일반 명사, 고유 명사, 관형사, 의존명사, 품사 등이었다. 특별히 숫자를 추가한 이유는 현재 상당수의 게임들이 "레시피"라 불리는 아이템 생성 공식이 존재하기에 숫자로도 어떤 게임을 플레이하는지 유추할 수 있기 때문이었다. 학습 알고리즘은 CBOW와 Skip-gram 알고리즘을 각각 사용하였으며 윈도우 사이즈를 5와 12로 하여 각각의 차이도 살펴보았다

4.2.1 Skip-gram

Skip-gram 결과를 먼저 살펴보면 연산 전 인접단어들의 결과는 [표 6], [표 7]과 같으며 품사 분리를 하지 않은 기존 CBOW 결과값보다도 상당히 떨어짐을 확인할 수 있었다. 또한 "인형-소전+칸무스" 테스트에서도 칸코레가 아닌 "벽람항로" 집합군을 찾는 모습을 보여주었다.

[표 6] 품사 분리 후 Skip-gram 알고리즘 및 윈도우 사이즈 5로 설정한 결과

[Table 6] Result of Skip-gram Algorithm with Window size 5 and Separated Sentence

단어	근접 단어
게임	사쿠라기, 신판, 우키, 증강현실, 여탕, 디벨로, 만서, 엎치락뒤치락, 디레
이벤트	브륀, 후편, 힐드, 나팔꽃, 팍스, 홍안, 연말연시, 날자, 총출동
소전	근사치, 섭이, 질럿, 럼, 모항, 네요, 후유코, 린데, 3550, 동기둥
소녀전선	
벽람	어흠, 근사치, 3550, 나침반, 소창, 랩소디, 연합국, 녹스, 항전, 보톰즈, 투항, 섭이
벽람항로	
깡겜	
칸코레	근사치, 후유코, 1670, 야도, 레이테, 게센, 난도, 아류, 사세보, 고니시, 똥칠, 깡창, 영전, 마나님, 장갑기병, 우군, 탐조등, 알콧, 에싯, 게임, 연주회, 시즈마, 연합국
재밌다	
재밌는	
시발	씨발, ㅋㅋ, 진짜, ㅋㅋㅋ, 아, 도끼질, 왈도체, 왜, 똥침, 망전, 거, 존나, 해달
씨발	시발, 진짜, 지금, ㅋㅋㅋ, ㅋㅋ, 왜, 다, 노량, 이, 또, 아, 다의, 거, 급방
소울워커	
인형-소전+칸무스	보톰즈, 깡, 근사치, 전소, 벽람, 화전민, 나침반, 똥칠, 배박이, 아케이드, 아류, 유저

[표 7] 품사 분리 후 Skip-gram 알고리즘 및 윈도우 사이즈 12로 설정한 결과

[Table 7] Result of Skip-gram Algorithm with Window size 12 and Separated Sentence

단어	근접 단어
게임	거, 사람, 안, 개인차, 생각, 금유, 자유도, 급부상, 데, 브로치, 잘
이벤트	힐드, 막간, 완수, 접두사, 할리, 성총, 곡집, 소절, 셜리, 프록시, 돌밑
소전	3550, 근사치, 180130, 장갑기병, 병종, 전희, 완수, 구운몽, 시프카, 모래사장
소녀전선	
벽람	소전, 나침반, 3550, 디스토리어, 보톰즈, 벽, 칸코레, 이소카제 아사시오, 시즈마
벽람항로	
깡겜	
칸코레	나침반, 보톰즈, 아사시오, 겜, 중파, 레이테, 마이즈루, 고니시, 항전, 항모
재밌다	
재밌는	
시발	ㅋㅋ, 아, 아니, 존나, 진짜, ㅋㅋㅋ, 씨발, 또, 와, 다
씨발	또, 뭐, 아니, 시발, 왜, 거, 아, 진짜, 존나, 안
소울워커	
인형-소전+칸코레	벽람, 보톰즈, 나침반, 칸무스, 배박이, 항모, 타카오, 벽, 근사치, 중파, 항전

4.2.2 CBOW

대부분의 경우 CBOW보다 Skip-gram이 더 정확하다는 평가와 달리 한국어를 이용하고 140자로 제한된 SNS 데이터에서는 CBOW의 정확도가 더 높았으며 벡터 연산 질문에 정확성 역시 높은 것이 확인되었다. 이는 기존의 Skip-gram으로 시도한 실험들이 영문으로 적힌 위키나 장편 소설 등 영어권 문장과 장문의 글을 상대로 학습시켰기 때문에 나타난 현상으로 보인다.

[표 8] 품사 분리 후 CBOW 알고리즘 및 윈도우 사이즈 5로 설정한 결과

[Table 8] Result of CBOW Algorithm with Window size 5 and Separated Sentence

단어	근접 단어
게임	자리매김, 노안, 울티마, 서관, 플라이트, 루미네스, 에누마, 프론트엔드, 코나미
이벤트	레이테, 이벤, 후편, 해역, 특효, 사세보, 코니시
소전	벽람, 레식, 물거북, 함포, 인기투표, 나침반, 카리나
소녀전선	
벽람	나침반, 소전, 근사치, 연합국, 코레, 전소, 칸코레, 우군, 소창, 코니시, 시즈마
벽람항로	

깡겜	
칸코레	우군, 코레, 인트레피드, 나침반, 전소, 코니시, 해역, 벽람, 항전, 레이터, 항로, 칸무스, 제독, 이나즈마, 딜라이트, 중파, 보크사이트
재밌다	
재밌는	
시발	지르, 씨발, 먼디, 다매, ㅋㅋ, 빡종, 호성, 멋져, 와아아
씨발	시발, 폐철, 청웅, 팀킬, 으아아아악
소울워커	
인형-소전+칸무스	칸코레, 벽람, 데차

[표 9] 품사 분리 후 CBOW 알고리즘 및 윈도우 사이즈 12로 설정한 결과

[Table 9] Result of CBOW Algorithm with Window size 12 and Separated Sentence

단어	근접 단어
게임	시작, 겜, 레볼루션
이벤트	알찬, 애니버서리, 특효, 이번, 봄맞이, 시작, 레이테, 시제
소전	명일, 성총, 항, 아동절
소녀전선	
벽람	나침반, 소전, 칸코레, 근사치, 시즈마, 코니시
벽람항로	
깡겜	
칸코레	
재밌다	
재밌는	
시발	아니, 진짜, ㅋㅋ, 아, 존나, 왜, 씨발, 개, 와, ㅋㅋㅋ
씨발	시발, 존나, 아니, 진짜, 놈, 왜, 개, 또, 야, ㅋㅋ
소울워커	
인형-소전+칸무스	칸코레, 벽람, 나침반, 코니시

4.2.3 벡터 연산을 통한 사용자 평가

앞선 실험의 결과로 "시발" 등의 비속어와 함께 게임에 대한 평가 단어도 인접한다는 사실을 알 수 있었다. 따라서 SNS 데이터에서 나타나는 비속어와 함께 게임 콘텐츠에 대한 실험을 수행하였다. 각 알고리즘과 윈도우 사이즈는 [표 10]에 정리하였다. 대부분의 경우 벡터 연산을 통해 게임 사용자들 사이에서 이슈가 되는 현상, 비난, 관심사 등의 키워드를 추출할 수 있었으며 게

임 타이틀과 콘텐츠가 사용자들 사이에서 어떤 위치에 있는지를 짐작할 수 있는 유의미한 단어들을 발견할 수 있었다.

[표 10] 알고리즘과 윈도우 사이즈에 따른 벡터 연산의 결과

[Table 10] Result of Vector Representation using different algorithm and window size

벡터 연산	알고리즘	의도에 유의미한 단어(거리)
시발 - 시바후 + 씨발	CBOW 윈도우 사이즈 5	인트레피드 (0.8844945) 칸코레 (0.8652503) 히무라 (0.8649219) 배박이 (0.86454) 우군 (0.8616309) 씨바후 (0.8549138) 나침반 (0.85318) 에식스 (0.8525645) 다나카 (0.8485526) 벽람 (0.8481827) 딜라이트 (0.847798) 항모 (0.8471619
	CBOW 윈도우 사이즈 12	인트레피드 (0.7759485) 박이 (0.7357869) 오오이 (0.7200369) 병신 (0.7114039) 칸코레 (0.7106965) 주구 (0.710115) 히무라 (0.7090307) 고니시 (0.7032924)
	Skip-gram 윈도우 사이즈 5	칸코레 (0.9861597)
	Skip-gram 윈도우 사이즈 12	인트레피드 (0.990277) 요크타운 (0.9877263)
물거북 - 소전 + 코니시	CBOW 윈도우 사이즈 5	
	CBOW 윈도우 사이즈 12	코레 (0.7488267) 벽람 (0.7427329) 칸코레 (0.7393131)
	Skip-gram 윈도우 사이즈 5	콩고 (0.9747644) 코레 (0.9747054) 칸코레 (0.9743829) 역작 (0.9743737)
	Skip-gram 윈도우 사이즈 12	아사시오 (0.9888457)

재미 – 넷마블 + 페그오	CBOW 윈도우 사이즈 5	딜라이트 (0.7409291) 복주머니 (0.7325587)
	CBOW 윈도우 사이즈 12	딜라이트 (0.6117013) 오다 노부나가 (0.589932) 칼데아 (0.5898017) 길가메서쉬 (0.5695847) 딜라이트 (0.5605741) 과금 (0.5563481)
	Skip-gram 윈도우 사이즈 5	터닝 포인트 (0.9579074) 오감 (0.9574301) 반야 (0.956284) 개돼지 (0.9560528)
	Skip-gram 윈도우 사이즈 12	돈줄 (0.9766153) 성녀 (0.9749525)
벽람 – 아타고 + 칸코레	CBOW 윈도우 사이즈 5	이소카제 (0.8707529)
	CBOW 윈도우 사이즈 12	경순양함(0.7228838)
	Skip-gram 윈도우 사이즈 5	콩고(0.9565294)
	Skip-gram 윈도우 사이즈 12	프린츠 오이겐(0.9800412)

우리는 전체적으로 CBOW 알고리즘에서 효율적인 단어 제시가 높음을 알 수 있다. 앞서 언급한 것과 같이 한국어 SNS 데이터가 가지는 특징이라 여겨진다. 주목할 만한 것은 초기 예상대로 사용자들의 호감과 비난 모두 비속어와 가까운 거리를 지닌다는 점이다. 특히 원화가 시바후에 대한 불만이 "인트레피드"라는 캐릭터 때문인 것을 정확하게 답을 제시하는 것은 확인할 수 있었으며 특정 게임에 대한 유저들의 비난도 확인 할 수 있었다. 이는 비속어를 제외시켰고 Skip-gram만으로 조사하던 기존 연구와의 차이점으로 기계학습에서 비속어와 대상 언어의 중요성을 일깨워 준다.

ISSN: 2384—101X JNCIST

5. 결론 및 후속 연구

이상으로 SNS 데이터와 Word2Vec를 이용하여 사용자들 사이에서 어떤 의미를 갖고 있는지를 살펴보았다. 게이머들은 한가지 게임만 하는 것이 아닌 비슷한 게임들을 플레이하거나 이슈를 자세히 알고 있었으며 자신이 플레이하는 게임에 대해 즉각적인 반응을 보여주고 있음을 확인할 수 있었다. 특히 SNS는 사회적 이슈에도 즉각적인 반응을 보여주기에 실험 결과와 같이 한달 이내의 짧은 기간으로도 의미 있는 고객 분석과 운영 방안을 세우는데 큰 도움이 될 것이다. 또한 영어권에서 제작된 스크립트로만 Word2Vec을 수행한다면 한국어를 분석하는데 정확도가 떨어질 수 있다는 사실도 결과로 나타났다.

후속 연구로는 데이터 분량과 학습 알고리즘에 있어 효율적인 공식을 찾아보고자 한다. 그를 위해 상시적인 자동 학습에 모듈을 제작해 학습의 편리성을 강화해 인공지능에게 각 단어의 관계와 함께 관계 발생 시기도 학습시킬 것이다. 또한 단어의 의미에만 국한되지 않고 한국어의 문장 구조까지 학습시켜 비지도 학습 기반의 완전한 문장을 출력하는 채팅봇을 개발할 예정이다.

References

[1] Lee Jin-Kyung "A Comparative Analysis of Success and Failure Cases on Advertising Marketing by Using Twitter" Incheon National University. College of Business Administration, (2014)

[2] "Usage Analysis of Swearing Words on Web Board and Proposal of Problems Resolution Method" The Korea Contents Society 3(4), (2013):1-10.

[3] Phil-Sik Jang. "Study on Principal Sentiment Analysis of Social Data." Journal of the Korea Society of Computer and Information , 19.12 (2014): 49-56.

[4] Woo Sung Chang, Yoon Taek Moon. "Exploratory Study on SNS (Facebook and Twitter) Brand Personality and Brand Loyalty Focused on Moderating Effect of the Involvement." The Korean Journal of Advertising and Public Relations, 14.2 (2012): 359-387.

[5] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781(2013).

[6] Muhammod Omar "Topic Coherence Evaluation: A Machine Learning Perspective" Yeungnam University. Department of Information and Communication Engineering, (2017)

[7] Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." Journal of Machine Learning Research 12.Jul (2011): 2121-2159.

[8] Elman, Jeffrey L. "Finding structure in time." Cognitive science14.2 (1990): 179-211.

[9] Mikolov, Tomáš, et al. "Recurrent neural network based language model." Eleventh Annual Conference of the International Speech Communication Association. (2010).

[10] Kudo, Taku. "Mecab: Yet another part-of-speech and morphological analyzer." http://mecab. sourceforge. jp (2006).

[11] Saggionα, Horacio, and Adam Funk. "Interpreting SentiWordNet for opinion classification." Proceedings of the seventh conference on international language resources and evaluation LREC10. (2010).