워드 임베딩을 활용한 관용표현 인식 연구

박서윤○, 강예지, 강혜린, 장연지, 김한샘†

연세대학교 언어정보학협동과정, 언어정보연구원†

{seoyoon.park, yjkang5009, hyerink, yeonji3547, khss†}@yonsei.ac.kr

Korean Idiom Classification Using Word Embedding

Seo-Yoon Park○, Ye-Jee Kang, Hye-Rin Kang, Yeon-Ji Jang, Han-Saem Kim†

Interdisciplinary Graduate Program of Linguistics and Informatics, Yonsei University

요 약

우리가 쓰는 일상 언어 중에는 언어적 직관이 없는 사람은 의미 파악이 힘든 관용표현이 존재한다. 관용표현을 이해하기 위해서는 표현에 대한 형태적, 의미적 이해가 수반되어야 하기 때문이다. 기계도 마찬가지로 언어적 직관이 없기 때문에 관용표현에 대한 자연어 처리에는 어려움이 따른다. 특히 일반표현과 중의성 관계에 있는 관용표현의 특성이 고려되지 않은 채 문자적으로만 분석될 위험성이 높다. 본 연구에서는 '관용표현은 주변 문맥과의 관련성이 떨어진다'라는 가정을 중심으로 워드 임베딩을 활용한 관용표현과 일반표현에 대한 구분을 시도하였다. 실험은 4개 표현에 대해 이루어 졌으며 Skip-gram, Fasttext를 활용한 방법을 통해 관용표현은 주변 단어들과의 유사성이 떨어짐을 확인하였다.

주제어: 관용표현, 중의성 해소, 워드 임베딩, Fasttext

1. 서론

일상적으로 쓰이는 말 중에는 언어 직관이 있는 화자여야만 의미를 이해할 수 있는 관용표현이 있다. 가령 다음 문장의 경우 한국어 직관이 있는 화자는 어떤 것이 직설적인 용법인지, 관용적 용법인지를 한눈에 구분할 수 있지만 기계는 불가능하다.

나는 아이의 **말을** 주의깊게 **들어주었다**.
철수는 엄마 **말을 잘 듣는** 착한 아이다.

최근 컴퓨터와 인간의 상호작용을 위한 자연어 처리 기술이 다양하게 연구되고 있으나, 인간의 언어적 직관을 반영한 사례는 찾아보기 어렵다. 언어 직관이 절대적으로 필요한 관용표현 또한 자연어 처리를 하기 위해서는 많은 어려움이 수반된다. 가령 관용표현에 형태소 분석이나 구문 분석을 하게 될 경우, 관용표현의 형태적 특징은 물론 의미적 특징 역시 상실된다. 또한 관용표현에는 형태적, 의미적 정보를 감지할 수 있는 외현적 특징이 없기 때문에, 기계적으로 표현을 구분하는 것 역시 불가능하다.

이에 대해 본 연구에서는 워드 임베딩 기법을 사용하여 관용표현과 일반표현의 구분을 시도하고자 한다. 관용표현과 일반표현의 중의성이 해소될 경우 기계번역 분야에서의 정확도 향상은 물론 동형이의어 처리, 구문 분석 및 의미역 분석 이전 shallow parsing이 가능하다는 점에서 의의가 있다.

본 논문의 구성은 다음과 같다. 2장에서는 연구 수행의 바탕이 된 선행 연구 및 관련 연구를 서술하였고, 3장에서는 연구에 사용된 데이터와 방법론에 대해 기술하였다. 4장에서는 관용표현과 일반표현 구분을 위한 연구과정을 다루었고, 마지막 5장에서는 결론 및 향후 연구에 대해 논의한다.

2. 선행 연구

외국의 관용표현 추출 및 중의성 해소 연구는 관용표현의 언어적 특성 자체에 주목한 Type-based 추출(extraction), 관용표현과 문맥과의 유사성(similarity)을 고려한 Token-based 추출로 나누어져 진행되었다.

Type-based 추출은 관용표현은 일반표현(literal expression)에는 존재하지 않는 언어적 특성이 있음을 전제로 한다[1]. 그 특성들은 세 가지로 어휘 교체 시 관용적 의미가 소실되는 어휘적 제약(lexical fixedness), 통사적 변화 시 관용적 의미가 소실되는 통사적 제약(syntactic fixedness), 그리고 제3의 의미를 가지는 비합성성(non-compositionality) 이다. Type-based 추출을 위해 어휘적 제약, 통사적 제약의 고정성(fixedness) 측정 방법인 PMI(Pointwise Mutual Information), 공기하는 벡터를 중심으로 관용표현과 일반표현을 구별하는 'Cform(Canonical Form)', 'Differ'과 같은 방법들이 사용되었다[2, 3].

Token-based 추출은 관용표현 추출 시 문맥과 token과의 관계성을 고려하는 방법이다. 이는 type-based 추출과정에서 문맥을 고려할 수 없었던 단점을 보완한 방법으로, 일반 표현은 주변 문맥과 강한 응집성(cohesion)을 가지는 반면 관용표현은 주변 문맥과 응집성이 떨어지는 것을 전제로 한다. Token-based 추출의 대표적인 방법론으로는 'Cohesion graph'가 있다[4, 5]. Cohesion graph에서 각 꼭지점은 분류하고자 하는 표현이 포함된 문장의 토큰들로, 각 토큰들은 문맥 안 다른 토큰들과의 의미적 관련성에 따라 가중치가 부여된 상태이다. 관용표현이 Cohesion graph에 표현된 경우 주변 문맥과의 유사성(similarity)이 떨어진다. 즉 관용표현은 주변 문맥과 응집성이 적은 표현이며, 이는 관용표현과 일반표현을 구분할 수 있는 기준이 되었다. 본 연구에서도 이를 전제로 삼아 워드 임베딩을 활용하여 탐지하고자 하는 표현과

문맥의 관련성을 탐지의 기준으로 삼았다.

2013년 워드 임베딩(Word Embedding [6, 7])이 고안된 이후에는 이를 이용하여 관용표현의 중의성 해소를 시도하는 연구들이 진행되었다[8-10]. Word2Vec을 활용한 방법은 type-based 추출처럼 언어의 특성을 고려할 필요도 없을 뿐 아니라, token-based 추출처럼 단어의 vector와 문맥(context)의 관계를 고려하여 관용표현과 일반표현의 분류를 진행할 수 있다. [8]의 경우 중의성을 가진 관용표현 중 빈도 수 상위 20위 단어를 대상으로 word2vec을 사용하여 일반표현과 구분하는 실험을 진행하였다. 2013년 [6, 7]에 의해 고안된 Word2Vec 이후에도 GloVe, Fasttext 등 단어를 벡터로 표현하는 다양한 방법들이 연구되었으며, 본 연구에서는 형태학적(morphological) 특성을 벡터 값에 반영할 수 있는 Fasttext를 사용하였다.

한국어의 어구추출은 주로 연어(collocation)를 대상으로 이루어졌다. 연어는 '두 개 이상의 단어가 결합하여 의미적으로 하나의 단위를 이루는 말'이다. '두 개 이상의 단어가 결합'하기 때문에 형태·통사적 제약성이 발생하며, 때문에 연어는 연어핵(node)과 연어변(collocate)로 구성되며 대체로 서로에 대한 공기성이 높다. 따라서 연어 추출 시도는 인접성, 공기성을 기준으로 연구되었으며, '계량언어학적 연구'라는 이름 아래 주로 진행되었다. 이러한 측면에서 이루어진 연구에는 [11-13] 등이 있다. 특히[13]에서는 연어핵과 한 문장에서 공기하는 연어변의 관계를 T검증, 포아송 분포, 인자분석을 활용하여 인접 공기관계, 구문적 공기관계, 군집 공기관계로 다각적으로 분석하였다.

자연어 처리 분야에서는 연어에 대한 '자동추출'이란 이름 아래 [14, 15]등의 연구가 진행되었다. [14]의 경우 연어를 추출하기 위해 중심어-연어변 간 거리 정보와 통계 정보를 이용하였다. 또한 선택적 제약 조건을 추출조건으로 설정하여 20만 어절 규모 말뭉치에서 3000개의 연어에 대해 약 90%의 정확도로 연어를 추출하였다[15]의 경우 동사 명사 조사 공기쌍을 대상으로 정규분포를 활용해 Z-검증 방법을 활용하였다. 이를 통해 동사에 대한 명사구의 분포가 우연으로 기대되는 것 이상으로 발생할 경우 이를 연어로 판단하였다.

관용표현의 경우 연어와의 형태·통사적 유사성을 공유한다. 관용표현은 연어와 마찬가지로 형태·통사적 제약성 및 공기제약성을 가지기 때문이다. 이에 따라 관용표현은 넓은 의미의 연어에 포함되며, 이를 일반표현과 구별하는 시도 역시 '연어' 범주에 포함되어 이루어졌다[12, 16, 17].

다만 관용표현은 연어와 달리 단어들을 통해 유추될 수 없는 제 3의 의미를 가진다. 즉 관용표현은 관용적 의미를 가지며, 관용표현에는 축자적 의미와 관용적 의미 두 개가 공존하는, 동음이의적 특성이 존재한다. 이 같은 특성은 관용표현으로 하여금 중의성(ambiguity)을 갖게 한다. 관용표현과 일반표현의 중의성 해소는 주로 격틀, 논항 자질, 의미자질과 같이 관용표현의 통사적, 의미적 특징을 일반표현과 비교하는 것을 통해 시도되었다[18-20].

한국어에서 딥러닝 기법을 활용한 중의성 해소는 주로 단어 단위로 시도되었다. [21]에서는 단어의 중의성을 해소하기 위해 Word2Vec을 활용한 임베딩 값을 이용해 cohesion graph를 사용하였고, [22]에서는 임베딩 단계에서 의미정보를 부착해 중의성을 해소하고자 하였다. 이처럼 한국어에서도 워드 임베딩을 통해 중의성을 해소하고자 하는 시도가 늘고 있다.

이처럼 국내외적으로 관용표현과 일반표현을 구분하고자 하는 시도는 관용표현의 언어적 특성을 반영한 통계적 방법론을 시작으로 관용표현과 문맥의 관계성, 그리고 딥러닝 기법을 활용하여 구분하는 방향으로 나아가고 있다.

3. 데이터 및 방법론

3.1. 관용표현 정의

관용표현이란 '두 개 이상의 단어로 이루어져 있으면서, 그 단어들의 의미만으로는 전체의 의미를 알 수 없는 표현'이다. 즉, 둘 이상의 말이 하나의 단위를 이루면서, 동시에 제 3의 의미를 가지는 어구이다. 제 3의 의미를 가지는 동시에 관용표현은 일반표현과 동음이의적 관계를 가지기 때문에, 중의성을 띄게 된다. 가령 '말을 듣다'의 경우 관용적으로 '꾸지람, 시비의 대상이 되다.', '기계, 도구 따위가 사람의 뜻대로 움직이다.'의 의미를 가지지만, 표면적으로는 '다른 사람의 발화를 귀로 듣다.'라는 의미를 가진다.

본 연구에서는 이처럼 관용표현을 '두 개의 단어로 구성되어, 단어의 뜻만으로는 유추될 수 없는 제 3의 의미를 가지는 표현'으로 정의하였다. 때문에 실험대상을 선정할 때 의미적으로는 일반표현과 중의성을 가지는지, '체언+용언'의 형태를 가지는지를 기준으로 하여 표현을 선정하였다. 실험 대상 관용표현을 체언+용언의 형태를 가진 표현으로 선정한 이유는 관용표현 중 '체언+용언' 형태가 가장 보편적인 빈도를 보이는 형태이기 때문이다[18]. 영미권의 경우에도 관용표현을 자동으로 탐지하는 연구가 Verb+Noun 꼴을 기본으로 하여 활발하게 이루어지고 있다[9, 10, 23, 24].

3.2. 방법론

본 연구에서는 관용표현과 일반표현의 단어 벡터를 활용하여 두 표현 간 구분을 시도하였다. 이를 위해 기존 연구 [8]에서 사용된 Skip-gram과, Fasttext 두 모델을 모두 실험에 활용하였다.

임베딩 값을 계산하는 가장 보편적인 방법으로는 Word2Vec이 있는데, Word2Vec은 주변 문맥을 통해 타겟 단어를 예측하는 CBOW와, 특정 단어를 통해 주변 문맥을 예측하는 Skip-gram으로 나누어진다. Word2Vec은 '단어'를 기준으로 하여 벡터 값을 만들어 낸다. 단어를 벡터화할 경우 단어의 의미를 여러 차원에 분산할 수 있으며, Word2Vec을 통해 학습된 벡터는 다른 자연어 처리 모델들의 입력 값으로 활용된다. 다만 Word2Vec은

단어를 최소 단위로 보기 때문에 학습하지 않은 단어에 대해 벡터 값을 만들어낼 수 없으며, 빈도 수가 낮은 표현에 대해 정확도 높은 임베딩을 만들어낼 수 없다.

본 연구의 중심인 Fasttext의 경우 단어 철자에 대한 n-gram을 학습한다. 즉 단어 안에 subword가 있는 것을 가정하기 때문에 Word2Vec과 달리 학습하지 않은 데이터에 대해서도 임베딩 값을 만들어낼 수 있다. 또한 데이터의 노이즈에 강하기 때문에, 데이터 전처리가 미진하더라도 정확한 벡터 값을 얻을 수 있다[25].

본 연구에서 Fasttext를 실험에 사용한 이유도 희소한 단어 및 훈련되지 않은 단어들에 대해서도 임베딩 값을 얻을 수 있고, 오타나 전처리가 미진한 실제 언어 자료에 효용성이 높은 모델이라 판단하였기 때문이다. 실제로 실험 결과 훈련하지 않은 데이터에 대해서도 유의미한 벡터 값을 얻을 수 있었다.

3.3. 데이터 선정 및 전처리

모델 훈련을 위한 훈련 데이터로는 세종계획 말뭉치(총3600만 어절) 중 무작위로 선정한 6,274,893어절, 연세 균형 말뭉치에서 '눈을 감다, 손을 들다, 말을 듣다, 물을 먹다'를 포함한 816,076어절을 선정하여 총 7,090,969어절 규모의 말뭉치를 훈련 데이터로 사용하였다. 훈련되지 않아도 단어의 벡터 값을 도출하는 Fasttext의 특성 상, 테스트를 위해 훈련 데이터에 포함되지 않은 20문장을 테스트 데이터로 사용하였다. 다만 이 20문장은 Skip-gram을 훈련할 때에 학습 데이터에 포함되었다.

테스트 데이터에는 위에 서술한 4개의 표현에 대해 관용표현으로 사용된 10문장, 일반표현으로 사용된 10문장이 포함되었다. 테스트 데이터는 무작위로 세종계획말뭉치, 연세균형말뭉치에서 추출하였다. 말뭉치 전처리로는 말뭉치에 부착된 메타정보 삭제, 문장단위 분절을 진행했으며, Skip-gram, Fasttext 각 모델에서는 원시 말뭉치에 대해 어절 별 토큰화, 문장 부호 삭제 등의 전처리를 하여 훈련을 진행하였다.

4. 실험

4.1. 실험대상

실험대상으로는 일반표현과 중의성을 가진 '체언+용언' 꼴 관용표현 4개를 선정하였다. 선정을 위해 각 사전[1]에서 수집한 관용표현 목록 중 상위 10위의 빈도를 보이는 명사를 포함한 표현으로 선정하였다.

표 1 관용표현 빈도 수 상위 10위 명사

1	눈(599)	6	머리(308)
2	말(499)	7	발(256)
3	손(459)	8	개(239)
4	입(396)	9	뒤(220)
5	물(315)	10	속(210)

*괄호 안은 빈도 수

관용표현의 경우 신체어의 빈도가 비신체어에 비해 높은 편이므로, 균형을 위해 상위 신체어 2개(눈, 손), 상위 비신체어(말, 물) 2개를 선정하였다. 이에 따라 선정된 표현은 '눈을 감다, 손을 들다, 말을 듣다, 물을 먹다'이며, 이번 연구에서는 각 표현과 문맥의 관계성을 명확하게 관찰하기 위해 체언과 용언 사이에 다른 단어가 첨입되지 않은 예들을 데이터로 활용하였다. 각 표현에 대한 일반표현과 관용표현의 의미는 다음과 같다.

눈을 감다
일반표현) 눈꺼풀을 내려 눈동자를 덮다.
관용표현) 남의 허물 따위를 보고도 못 본 체하다.

손을 들다
일반표현) 신체기관 손을 위로 올리다.
관용표현) 자기 능력에서 벗어나 그만두다.

말을 듣다
일반표현) 다른 사람의 말이나 소리에 스스로 귀 기울이다.
관용표현) 꾸지람을 듣거나 시비의 대상이 되다,
기계·도구 따위가 다루는 사람의 뜻대로 움직이다.

물을 먹다
일반표현) 음식 따위를 입을 통하여 배 속에 들여보내다.
관용표현) 어떤 것을 따르거나 영향을 받다.

특히 일반표현, 관용표현 각 10문장으로 이루어진 테스트 데이터의 경우 위에 제시된 의미를 기준으로 선정하였으며, 각 테스트 문장은 최대 길이가 20어절이 되도록 데이터를 구성하였다. 테스트 데이터 예시는 다음과 같다.

종류	예문		
관용	하여튼 서울	물을 먹어본	내 눈에는 시골 장날의 축제 기분이라는 것이... 틀림없겠지만
일반	음식물이나	물을 먹지	않고 살 수 있는 실험 ...세균을 주사하는 실험하기도 하였다
관용	그러자면 검문소와 경찰 군부대까지	눈을 감아	주도록 손을 써야 했고 신속하게 해치워야만 뒤탈이 없었다
일반	돌아오는 도중 소용이는 시종일관	눈을 감고	한 마디의 말도 없었다 21일 동안 떠나 있던 집에 돌아왔다
관용	먼저 미국과 독일 영국의 거상들이	손을 들고	말았다 그들 나라에서 ... 치명상이 되었다
일반	k씨의 경우 ys와 함께	손을 들고	있는 사진은 ... 서울시장 사진을 더 크게 썼다
관용	그러면 농협 간부진이 농림부의	말을 듣지	않으면 되지 않느냐고 하지만 그것이 어렵다는 것이다
일반	것은 닭장 위에다 토끼를 기르면 좋다는	말을 듣고	한 마리 얻어 오면서부터입니다

그림 1 테스트 데이터 예시

4.2. 실험방법

Skip-gram과 Fasttext로 워드 임베딩을 진행하기 위해 오픈소스 라이브러리 Gensim을 사용하였다. 두 모델은 모두 Gensim에 구현되어 있으며, 앞서 언급한 바와 같이 모델의 훈련을 위해 약 7백만 어절의 말뭉치를 학습하였다. Fasttext의 경우 파라미터는 학습률 0.05, 차원수는 100차원으로 각각 설정하였다. 윈도우 크기의 경우 중심 단어의 문자(character)를 10개씩 참고하도록 10으로 설정하였는데, 이는 각 문장이 20어절 이상으로 이루

[1] 한국어 기본 숙어 사전, 노용균, 한국문화사 (2002),
관용어 사전, 박영준 최경봉, 태학사(1996),
표준국어대사전, 국립국어연구원, 두산동아(1999),
연세한국어사전, 연세대 언어정보개발연구원, 동아출판사(1998),
우리말 큰사전, 한글학회, 어문각(1991)

어진 훈련데이터를 효율적으로 학습하고자 설정한 크기이다. 총 epoch 수는 5로 설정하였다.

Fasttext뿐만 아니라 Skip-gram을 활용한 이유는 기존연구 [8]에서 쓰였을 뿐만 아니라 Fasttext에 대한 실험의 대조군으로 활용하기 위함이다. 실험 조건을 맞추기 위해 window 크기와 차원의 수를 각각 10, 100으로 설정하였으며, iteration 수 역시 Fasttext와 동일하게 5로 설정하였다.

관용표현과 주변 단어(local text)의 긴밀성을 측정하기 위해 [8]에서 제시한 방법을 활용하였다. [8]에서는 local context와 중심표현 모두에 대해 벡터 값을 얻은 후, 이 둘을 내적(*inner product*)하여 상관성을 구하였다. V+N 형태의 idiom에 대한 벡터 값 v, n을 구하기 위해 두 단어에 대한 벡터 값을 각각 구하고, 이를 더하여 idiom에 대한 벡터 σ_{vn}로 상정하였다(1). 또한 해당 표현 주변의 m개 단어에 대한 벡터 값 v_m을(2) 구해 σ_{vn}과 곱한 후, 단어 m개에 대한 평균 p을 구하였다(3).

$$\sigma_{vn} = v + n \in R^q \quad (1)$$

$$V = [v_1, v_2, \dots v_m] \in R^{q \times m} \quad (2)$$

$$P = V^t \sigma_{vn} \quad (3)$$

4.3. 실험결과

4.2. 에서 기술한 바와 같이 본 연구에서도 테스트 데이터에 대해 중심 표현에 대한 체언(n), 용언(v)에 대한 벡터의 합(σ_{vn})을 중심 표현에 대한 벡터 값으로 사용하였다. 다만 어절을 단위로 하여 모델을 훈련시켰으므로 테스트 역시 어절 중심으로 진행하였고, 때문에 중심 표현에 대한 벡터 값을 얻을 때 체언의 조사 교체, 용언의 어미 교체를 고려하지 않고 한 단위로 처리하였다.

실험은 Skip-gram과 Fasttext를 사용하여 진행하였고, 각 모델에서 훈련을 통해 얻은 중심표현의 벡터 값 σ_{vn}을 문장 내 단어들의 벡터 값 $v_1, v_2, \dots v_m$ 과 내적한 후 단어 수에 m개에 대한 평균 w (Skip-gram의 결과), f(Fasttext의 결과) 를 구하였다. 4개 표현 각각에 대해 일반표현 10개, 관용표현 10개의 w, f값을 각각 구하였으며, 결과는 다음과 같다.

표2 Skip-gram 실험 결과

말을 듣다			물을 먹다		
	관용	일반		관용	일반
1	17.34	18.67	1	7.05	9.79
2	7.20	8.55	2	6.74	9.14
3	11.52	13.03	3	7.96	7.89
4	8.46	10.50	4	8.19	8.40
5	7.20	8.84	5	14.09	8.83
6	11.07	10.58	6	9.28	9.90
7	13.08	8.65	7	7.03	5.78
8	7.82	14.15	8	3.43	11.81
9	9.10	9.50	9	8.37	9.84
10	6.57	15.91	10	7.44	12.81
평균(*w*)	**9.94**	**11.84**	평균(*w*)	**8.26**	**9.42**
손을 들다			눈을 감다		
	관용	일반		관용	일반
1	9.14	8.85	1	9.73	12.53
2	5.93	11.85	2	13.67	14.70
3	10.87	13.34	3	10.11	11.10
4	8.64	10.22	4	8.69	16.77
5	11.30	14.51	5	15.60	11.48
6	14.31	13.12	6	12.66	11.95
7	9.78	6.93	7	8.92	17.75
8	9.91	9.01	8	8.59	17.49
9	7.75	11.94	9	13.27	20.97
10	7.58	13.68	10	10.36	13.35
평균(*w*)	**9.52**	**11.34**	평균(*w*)	**11.16**	**14.81**

표3 Fasttext 실험 결과

말을 듣다			물을 먹다		
	관용	일반		관용	일반
1	7.54	8.2	1	6.9	9.11
2	4.69	6.41	2	4.93	9.87
3	5.83	6.94	3	6.06	11.13
4	4.19	6.01	4	6.36	5.67
5	6.3	5.95	5	8.73	7.33
6	5.96	6.64	6	6.41	4.65
7	6.45	7.76	7	6.08	7.03
8	6.05	6.08	8	4.74	8.28
9	6.57	5.83	9	5.75	6.18
10	5.5	6.8	10	5.1	7.14
평균(*f*)	**5.91**	**6.66**	평균(*f*)	**6.11**	**7.64**
손을 들다			눈을 감다		
	관용	일반		관용	일반
1	5.96	4.78	1	5.32	6.32
2	3.81	7.75	2	6.04	7.03
3	5.5	7.67	3	5.61	5.34
4	3.9	5.06	4	5.51	7.49
5	7.22	6.75	5	9.21	6.37
6	6.46	8.51	6	5.4	6.75
7	5.78	5.51	7	4.59	7.2
8	5.11	6.35	8	6.87	8.25
9	6.04	5.25	9	5.95	8.83
10	5.2	7.76	10	6.77	8.06

평균(f)	5.5	6.54	평균(f)	6.13	7.16

총 80개에 문장에 대한 실험 결과들 중 일반표현으로 쓰인 경우가 관용표현으로 쓰인 경우보다 w,f값이 대부분 높게 나타났다. 따라서 일반표현은 주변 단어들과 관련성이 높고 관용표현은 반대인 것을 알 수 있다. 이는 '관용표현은 주변 문맥과 낮은 관련성을 가진다'라는 전제를 뒷받침한다. 또한 전체에 대한 평균을 살펴볼 경우 관용표현으로 쓰였을 때의 결과가 일반표현으로 쓰였을 때보다 모두 낮아, 관용표현을 탐지할 때에 주변 문맥과 중심 표현과의 관련성이 주요한 구분점임을 확인할 수 있었다.

Skip-gram과 Fasttext 실험의 관용표현에 대한 정밀도(Precision), 재현율(recall), 정확도(Accuracy)는 다음 표와 같다.

표4 Skip-gram의 분류성능평가지표

Skip-gram	Precision	Recall	Accuracy
말을 듣다	66.67	80	70
물을 먹다	81.82	90	85
손을 들다	75	90	80
눈을 감다	90	90	90
전체	77.78	87.5	81.25

표5 Fasttext의 분류성능평가지표

Fasttext	Precision	Recall	Accuracy
말을 듣다	90	90	90
물을 먹다	81.82	90	85
손을 들다	75	90	80
눈을 감다	90	90	90
전체	83.72	90	86.25

위와 같이 Skip-gram보다 Fasttext를 사용하여 관용표현을 탐지했을 때에 정밀도, 및 재현율, 정확도 모두 향상된 것을 알 수 있다. 이로 미루어보아 조사, 어미가 발달한 한국어를 임베딩 할 경우 단어 철자에 대해 n-gram 학습을 하는 Fasttext가 성능이 좋음을 확인할 수 있었다.

5. 결론 및 향후 연구

본 연구에서는 서로 중의성 관계에 놓인 관용표현과 일반표현을 구분하기 위하여,'관용표현은 주변 문맥과 관련성이 떨어진다.'라는 전제 아래 각 표현이 문장 내에서 주변 단어들과 얼마나 유사성을 나타내는지를 측정하였다. 측정을 위해서 중심 표현 벡터 값과 주변 단어들의 벡터 값을 내적한 합을 단어 개수로 나누어 평균을 구하였다. 실험 결과 관용표현은 내적 평균 값이 일반표현보다 작음을 확인하여 문맥 단어들과 유사성이 떨어짐을 확인하였다. 또한 일반표현과 관용표현을 구분할 때에 문맥과의 관련성이 주요한 기준이 됨을 확인하였다.

다만 이번 연구의 경우 체언+용언 사이에 아무런 단어도 들어가지 않은 기본 형태로만 실험하였고, 또한 가중치를 고려하지 않아 실험 결과에서 부분적인 예외가 발생하였다. 향후 연구에서는 부정표현, 수식어 등 중심 표현 안에 다른 단어들이 들어갔을 때 중심 표현의 벡터 값을 얻는 방법을 연구할 예정이다. 또한 관용표현과 일반표현을 정확하게 구분할 수 있는 최적의 가중치 부여 방법을 찾아 이를 고려한 연구를 진행할 예정이다.

참고문헌

[1] Ivan A. Sag et al. Multiword Expression: A Pain in the Neck for NLP, 2002

[2] Fazly, Afsaneh, and Suzanne Stevenson. "Automatically constructing a lexicon of verb phrase idiomatic combinations." 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006

[3] Paul Cook, Afsneh Fazly, Suzanne Stevenson, Pulling their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context, A Broader Perspective on Multiword Expressions p.41-48, 2007

[4] Caroline Sporleder, Linlin Li, Unsupervised Recognition of Literal and Non=Literal Use of Idiomatic Expressions, Conference of the European Chapter of the ACL p.754-762, 2009

[5] Hessel Haagsma, Malvina Nissim, Johan Bos, The other Side of the Coin: Unsupervised Disambiguation of Potentially Idiomatic Expressions by Contrasting Senses, Proceedings of the Joint Workshop on LAW-MWE-CxG-2018, p.178-184, 2018

[6] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space, IN: Proceddings of Workshop at ICLR, 2013

[7] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceddings of NIPS, 2013

[8] Jing Peng, Anna Feldman, Hamza Jazmati, Classifying Idiomatic and Literal Expressions Using Vector Space Representations, Proceedings of Recent Advances in Natural Language Processing(RANLP), p.507-

511, 2015

[9] Jing, Peng, Anna Feldman, Automatic Idiom Recognition with Word Embeddings, 2017

[10] Rafael Ehren, Literal of idiomatic? Identifying the reading of single occurences of German multiword expressions using word embeddings, 2017

[11] 홍종선, 강범모, 최호철, 한국어 연어 정보의 분석 응용에 대한 연구, 한국어학 11 p.73-158, 2000

[12] 서상규, 한국어 정보 처리와 연어 정보, 국어학 39 p.321-353, 2001

[13] 박병선, 2003, 국어 공기관계의 계량언어학적 연구, 고려대학교 박사학위논문, 2003

[15] 윤준태, 구문 분석을 위한 말뭉치부터의 어휘정보 획득 및 응용, 연세대학교 언어정보연구원 학술발표 논문집 창간호, 1998

[14] 이공주, 김재훈, 김길창, 품사 태깅된 말뭉치로부터 한국어 연어 추출, 한국정보과학회 학술발표논문집 제22권 제2호(A), p.623-626, 1995

[16] 한영균, 명사+동사' 합성구의 형태론적 특성- <동사 합성구 사전>의 거시구조와 관련된 문제를 중심으로, 울산어문논집 12, p.95-123, 1997

[17] 김진해, 관용어의 직설의미와 관용의미의 관계 연구, 한국어 의미학 13권0호, p.23-41, 2003

[18] 김한샘, 현대국어 관용구의 계량언어학적 연구, 연세대학교 석사학위논문, 1999

[19] 박세영, 국어 관용구 판정에 대한 연구, 고려대학교 석사학위논문, 2001

[20] 권경일. 국어 관용구 연구, 연세대학교 박사학위논문, 2005

[21] 오동석, 강상우, 서정연, Word2Vec을 이용한 반복적 접근 방식의 그래프 기반 단어 중의성 해소. 인지과학, 27(1), p.43-60, 2016

[22] 김홍진, 김학수, 딥러닝을 이용한 한국어 어의 중의성 해소, 제31회 한글 및 한국어 정보처리 학술대회 논문집, p.380-382, 2019

[23] Afsaneh Fazly, Paul Cook, Suzanne Stevenson, Unsupervised Type and Token Identification of Idiomatic Expressions, Association for Computationl Linguistics, 2009

[24] Agnes Tutin, Emmanuelle Esperanca-Rodier, The Difficult Identification of Multiword Expressions: From Decision Criteria to Annotated Corpora, Computational and Corpus-Based Phraseology (pp.404-416), 2019

[25] Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov, Enriching Word Vectors with Subword information, 2016