한국어 다의어 의미 구분을 위한 문맥 기반 클러스터링 임베딩 기법 연구

강다현, 김혜진, 박승현*
한성대학교 컴퓨터공학부
{ dhk02, 2271340, sp }@hansung.ac.kr

Context-based Clustering Embedding Method for Distinguishing Word Senses in Korean Polysemy

Dahyeon Kang, Hyejin Kim, Seunghyun Park*
Division of Computer Engineering, Hansung University

요 약

한국어 다의어 의미 구분의 정확성 향상을 위해 비지도 학습 방식인 클러스터링 기반 문맥 임베딩 기법을 제안하고 기존의 단일 평균 임베딩 방식과 성능을 비교하였다. 이를 위해 다의어를 포함하는 다양한 문맥 데이터를 수집하고, KoBERT를 이용해 임베딩 벡터를 추출하였다. 실루엣 점수와 이너셔를 통해 성능을 비교 평가한 결과, 클러스터링 기반 임베딩 방식이 두 평가 지표 모두에서 우수한 성능을 나타냈다. 제안 방법은 별도의 주석 없이 순수 비지도 학습만으로 문맥에 따른 의미를 명확히 표현할 수 있어 기계번역 등 다양한 자연어 처리 응용 분야의 기반 기술로 활용될 것으로 기대된다.

Ⅰ. 서론

최근 자연어 처리 기술은 문맥을 반영하여 단어 및 문장 의미를 보다 정교하게 임베딩하는 방향으로 발전하고 있다. 한국어는 외래어와 고유어가 혼합되어 사용되기 때문에 같은 형태의 단어라도 여러 가지 다른 의미를 포함할 수 있다. 이 같은 다의어가 정확하게 처리되지 않으면 의미 간 경계가 모호해져 자연어 처리 전반의 성능을 저하시킬 수 있으므로, 한국어에서 다의어 처리 연구의 중요성은 더욱 강조된다.

기존 단어 임베딩 방식은 대상 단어가 등장하는 각 문맥에서 임베딩 벡터를 추출한 뒤, 이를 단순히 평균하여 하나의 대표 벡터로 만드는 방식을 주로 사용한다. 그러나 단일 평균 임베딩은 서로 다른 문맥에서 나타나는 단어의 의미 차이를 제대로 반영하지 못하며, 특히 다의어의 여러 의미가 하나의 벡터로 혼합되어 표현력을 약화시키는 한계가 있다. Arora 등 [1]은 평균 임베딩 방식이 문맥의 다양성을 포착하지 못한다는 한계를 지적하였다. 이를 극복하기 위한 방안으로, 문맥 임베딩을 군집화하여 의미별 다중 벡터를 생성하는 기법이 제안되었다 [2,3].

그러나 한국어 다의어 처리와 관련된 선행 연구는 여전히 부족한 편이며 [4], 기존 연구 역시 제한적 어휘셋이나 규칙 기반 방법에 의존하는 경우가 많았다. 최근 다의어 주석 말뭉치를 활용한 지도 학습 기법이 일부 시도되었지만, 순수한 비지도 학습 방식만으로 한국어 다의어 의미를 구분하고 그 성능을 평가한 연구는 아직 미진하다.

이에 본 연구에서는 한국어 위키백과에서 다의어가 사용된 다양한 문맥을 수집하여, 단일 평균 임베딩과 클러스터링 기반 임베딩 방식을 비교 분석하였다. 특히, 클러스터링 방식의 대표적인 알고리즘인 K-Means를 이용해 문맥 벡터를 클러스터링하고, 각 군집의 중심값을 해당 단어의 서로 다른 의미를 나타내는 별도의 임베딩으로 활용하는 전략을 제안하였다. 이를 통해 평균화로 인한 의미 희석 문제를 완화하고, 다의어가 지닌 복수의 의미를 명확히 구분할 수 있는지 실험적으로 검증한다.

Ⅱ. 본론

1) 데이터 수집 및 전처리

본 연구는 한국어 다의어 의미 구분 능력을 실험적으로 검증하기 위해, 위키백과의 '자유' 항목에 포함된 문서를 중심으로 데이터를 수집하였다. 수집 과정에는 한국어 위키백과 API를 활용하였으며, 수집된 원문은 여러 단계의 전처리 과정을 통해 실험에 적합한 형태로 정제하였다. 분석 대상은 총 84개 문장으로, 다양한 의미로 사용된 '자유'라는 명사를 포함하고 있어 다의어 맥락을 탐색하기에 적합하다고 판단하였다. 명사 추출에는 통계적 접근 방식을 기반으로 한 Soynlp의 Noun Extractor ver. 2 알고리즘 [5]을 활용하였다. 불용어 처리를 위해 실질 자립 형태소 293개를 참고하였으며, 본 실험에서는 Stopwords-ISO 프로젝트 [6]의 한국어 불용어 리스트를 적용하였다. 이를 통해 의미 분석의 정밀도를 높이고, 문맥 임베딩 과정에서 불필요한 요소를 효과적으로 제거하였다. 단일 음절 명사는 분석 효율을 위해 제거하였으며, 이렇게 정제된 데이터는 이후 문맥 임베딩 실험의 입력으로 활용되었다.

2) 클러스터링 기반 임베딩

기존의 단일 평균 임베딩은 KoBERT 모델을 활용하여 각 문장에서 대상 명사가 등장하는 위치의 임베딩 벡터를 추출한 뒤, 이들 벡터를 전체 문맥에 걸쳐 평균하여 하나의 대표 벡터를 구성하는 방식이다. 이 방식은 구현이 비교적 간단하고 계산 효율성이 뛰어난 장점이 있다. 그러나 서로 다른 문맥에서 나타나는 다의적 의미가 단일 벡터로 통합됨으로써, 의미 간의 경계가 모호해지고 표현력이 저하된다는 한계가 존재한다.

이러한 문제를 극복하기 위해 본 연구에서는 클러스터링 기반 임베딩 방법을 제안한다. 클러스터링 기반 임베딩은 대상 명사가 사용된 문맥 벡터들을 군집화하여 서로 다른 의미를 개별적으로 표현하는 접근법이다. 본 연구에서는 대표적인 군집화 알고리즘인 K-Means를 활용하여 각 문맥 벡터를 클러스터링하였다. 실험 과정에서 클러스터 수를 사전에 고정하지 않고, 다양한 값을 실험적으로 적용하여 가장 적합한 군집 수를 도출하였다.

군집화가 완료된 후, 각 클러스터에 속한 문맥 벡터들의 중심값을 계산하여 이를 해당 클러스터의 대표 임베딩 벡터로 활용하였다. 예를 들어, '자유'라는 단어의 문맥 벡터를 두 개의 클러스터로 군집화할 경우, 정치적 권리로서의 '자유'와 개인적 심리 상태로서의 '자유'로 구분될 수 있으며, 각 의미에 대응하는 별도의 임베딩 벡터가 생성된다. 이와 같은 클러스터링 방식을 적용하면, 하나의 단어가 문맥에 따라 갖는 서로 다른 의미를 각기 다른 임베딩 벡터로 구분하여 표현할 수 있다. 결과적으로, 단일 평균 임베딩 방식과 비교하여 클러스터링 기반 임베딩이 문맥에 따른 의미 차이를 벡터 공간에서 더욱 명확하게 구분하고 표현할 수 있다.

3) 실험결과 분석

본 연구에서는 단일 평균 임베딩 방식과 클러스터링 기반 임베딩 방식의 성능을 정량적 평가와 정성적 평가로 나누어 분석하였다. 먼저, 정량적 평가는 대표적인 군집 품질 측정 지표인 실루엣 점수(Silhouette Score)와 이너셔(Inertia)를 활용하여 수행하였다. 실루엣 점수는 각 데이터 포인트가 속한 군집 내 응집도와 군집 간 분리도를 동시에 고려하여 계산하는 지표로, 값이 1에 가까울수록 군집이 명확하게 구분되었음을 의미한다. 이너셔는 각 데이터 포인트가 소속된 클러스터의 중심으로부터의 거리의 합을 나타내는 지표로, 값이 낮을수록 군집 내의 데이터들이 밀집되어 있음을 나타낸다.

실험 결과, 단일 평균 임베딩 방식의 경우 실루엣 점수는 0.0327, 이너셔는 8223.65로 측정되었으며, 클러스터링 기반 임베딩 방식은 10개의 클러스터를 기준으로 하여, 실루엣 점수 0.0427, 이너셔 6983.35로 나타났다. 이러한 결과는 클러스터링 기반 방식이 단일 평균 임베딩 방식에 비해 군집 내 응집도가 높고 군집 간 분산도가 낮아, 문맥에 따른 다의어의 의미를 보다 명확하게 구분할 수 있음을 시사한다.

표 1. 단일 임베딩과 클러스터링 임베딩 성능 비교

	단일 평균 임베딩	클러스터링 기반 임베딩 ($k = 10$)
실루엣 점수	0.0327	**0.0427**
이너셔	8,223.65	**6,973.35**

정성적 평가에서도 클러스터링 기반 임베딩 방식의 우수성을 확인할 수 있었다. 단일 평균 임베딩 방식의 경우 서로 다른 문맥에서 사용된 단어가 하나의 대표 벡터로 통합됨으로써 의미 간 경계가 모호해지고, 서로 다른 의미가 명확히 구분되지 않는 한계가 관찰되었다. 예컨대, '사람'이라는 명사는 철학적 맥락과 생물학적 맥락에서 사용된 문장들이 하나의 벡터에 병합되어 의미 간 구분이 어려웠다. 반면, 클러스터링 기반 임베딩 방식은 의미적으로 상이한 문맥을 서로 다른 클러스터로 구분하여 표현함으로써, 다의어가 지닌 실제 의미 차이를 보다 명확하게 나타냈다. 이는 클러스터링 기반의 임베딩 방법이 한국어 다의어 처리에서 의미 간의 경계를 보다 효과적으로 반영할 수 있음을 시사한다.

클러스터 수에 따라 군집 품질 변화를 실험적으로 분석한 결과, 클러스터링 기반 임베딩 방식이 전반적으로 더 낮은 이너셔 값과 개선된 실루엣 점수를 나타내었다. 이는 클러스터링 방식이 문맥의 유사성에 따라 보다 밀집된 군집 구조를 형성할 수 있음을 의미한다.

Ⅲ. 결론

본 연구에서는 한국어 다의어 의미 구분을 위한 비지도 임베딩 기법으로서, 단일 평균 임베딩과 클러스터링 기반 임베딩을 실험적으로 비교 평가하였다. 위키백과에 등장하는 '자유' 관련 문맥을 수집하여 KoBERT로 임베딩한 뒤, 평균 임베딩과 K-Means 기반 군집 임베딩의 성능을 실루엣 점수와 이너셔로 측정한 결과, 본 연구에서 제안하는 클러스터링 기반 임베딩이 의미 간 경계를 보다 명확히 반영함을 확인하였다. 또한 정성적 분석에서도 서로 다른 문맥을 제대로 구별해내는 모습을 보이며, 평균 임베딩 방식의 한계를 보완할 수 있음을 확인하였다.

본 연구의 결과는 문맥에 따른 다의어 의미의 차이를 벡터 공간 상에서 정밀하게 표현할 수 있음을 보여준다. 특히 별도의 수작업 주석 없이 순수 비지도 학습만으로도 효과적으로 다의어 의미를 구분할 수 있어, 실질적인 데이터 수집 및 분석 비용을 절감할 수 있다는 점에서 학술적, 실용적 가치가 높다고 평가할 수 있다. 제안한 방법은 기계번역, 정보 검색, 질의응답 시스템 등 다양한 자연어 처리 응용 분야에서 보다 정확하고 정교한 의미 표현을 지원하는 기반 기술로 활용될 수 있을 것으로 기대된다.

향후 연구에서는 보다 다양한 한국어 다의어 단어를 포함하는 대규모 말뭉치를 구축하여 연구 결과의 일반화 가능성을 높일 계획이다. 또한 문장 수준의 임베딩 기법을 적용하고, 지도 학습 기반의 Fine-tuned KoBERT 모델을 병행적으로 도입하여 의미 구분 성능을 더욱 향상시키고자 한다. 이는 기존의 비지도 임베딩 기법과 보완적으로 작용하여, 보다 정교하고 의미 분화된 표현을 가능하게 한다. 추가적으로 DBSCAN, HDBSCAN 등 다양한 군집화 알고리즘을 통해 군집 구조 및 최적 클러스터 수 결정 문제를 심도 있게 연구할 계획이며, 골드 스탠더드 말뭉치를 활용하여 실제 의미와 군집 결과 간의 정확한 일치도를 검증함으로써 연구의 신뢰성과 실용성을 지속적으로 확대할 계획이다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2022R1A2C2005705, 분산 머신러닝 기반 지능형 플라잉 기지국을 위한 AI-MAC 프로토콜).

참 고 문 헌

[1] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *Proc. of the 5th International Conference on Learning Representations (ICLR)*, Toulon, France, 2017.

[2] J. Reisinger and R. J. Mooney, "Multi-prototype vector-space models of word meaning," in *Proc. of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, Los Angeles, CA, USA, 2010, pp. 109–117.

[3] E. Huang, R. Socher, C. Manning, and A. Ng, "Improving word representations via global context and multiple word prototypes," in *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, Jeju, Republic of Korea, 2012, pp. 873–882.

[4] J.-C. Shin, J.-S. Lee, and C.-Y. Ock, "Korean polysemy word-sense disambiguation using MoDu-corpus," in Proc. of the 32nd Annual Conference on Hangeul and Korean Language Information Processing (HCLT), Seoul, Republic of Korea, 2020, pp. 205–208.

[5] H. Kim, "Noun Extractor ver 2," GitHub (online), https://github.com/lovit/soynlp, Accessed: 2025-05-07.

[6] Stopwords-ISO, "Korean Stopwords List," GitHub (online), https://github.com/stopwords-iso/stopwords-ko, Accessed: 2025-05-07.