고려대학교 KOREA UNIVERSITY

이상근 교수 (고려대학교 정보보호대학원)

자연어 처리 시스템 및 자연어 처리에서의 단어 표현 방법
(기술분류-보안-네트워크·클라우드 보안)

기술성 분석

기술 개요

- 본 기술은 기학습된 단어 임베딩의 지도 학습을 기반으로 미등록 단어(Out Of Vocabulary, OOV)를 비롯한 모든 단어에 대한 단어 표현을 생성하는 자연어 처리 시스템 및 자연어 처리에서의 단어 표현 방법에 관한 것임
- 미등록 단어뿐만 아니라 모든 단어의 하위단어정보를 이용하여 해당 단어가 가지고 있는 고유한 의미를 정확히 추출하고, 미등록 단어가 많은 개방 어휘 환경에서 효과적으로 동작할 수 있으며, 새로운 단어 임베딩을 생성하기 위해 말뭉치, 즉 대형 코퍼스에서 오랜 시간 학습하지 않고 기존의 자연어 처리 시스템이 가지고 있는 단어 임베딩을 이용하여 미등록 단어를 생성할 수 있기 때문에 단어 임베딩 생성에 있어서의 효율성 및 효과성이 향상될 수 있음

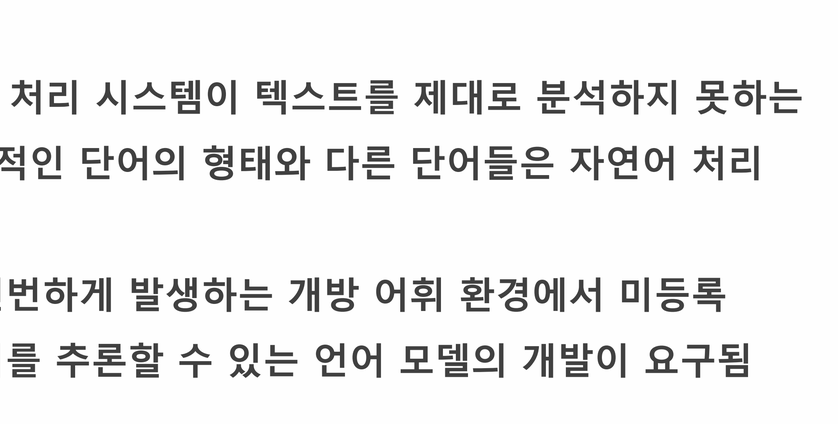
미해결 과제(Unmet needs)

- 기존 자연어 처리에서 미등록 단어 처리 방법의 한계
  - 기존의 미등록 단어의 처리 방법은 미등록 단어의 수가 많아질 경우에 자연어 처리 시스템이 텍스트를 제대로 분석하지 못하는 결과를 초래하는 문제점이 있으며, 특히 소셜 미디어 환경에서 사용되는 일반적인 단어의 형태와 다른 단어들은 자연어 처리 시스템이 지니고 있는 단어 임베딩에 존재하지 않을 확률이 매우 높음
  - 따라서, 자연어처리 시스템이 처리해야 할 단어의 수가 많거나 신조어 등이 빈번하게 발생하는 개방 어휘 환경에서 미등록 단어들 각각에 대한 표현을 생성하고, 미등록 단어가 지니고 있는 고유한 의미를 추론할 수 있는 언어 모델의 개발이 요구됨

기술적 해결수단(발명의 구성)

- 1) 본 발명에 따른 자연어 처리 시스템의 구성
  - 본 발명의 자연어 처리 시스템(100)은 통신 모듈(110), 메모리(120), 프로세서(130) 및 데이터베이스(140)로 구성됨
  - 통신 모듈은 통신망과 연동하여 자연어 처리 시스템과 사용자 단말 간의 송수신 신호를 패킷 데이터 형태로 제공하는 데 필요한 통신 인터페이스를 제공함
  - 메모리는 자연어 처리에서의 단어 표현 방법을 수행하기 위한 프로그램이 기록되며, 프로세서가 처리하는 데이터를 일시적 또는 영구적으로 저장함
  - 프로세서는 자연어 처리에서의 단어 표현 방법을 제공하는 전체 과정을 제어함
  - 데이터베이스에는 자연어 처리에서의 단어 표현 방법을 수행하면서 누적되는 데이터가 저장됨

본 발명에 따른 자연어 처리 시스템

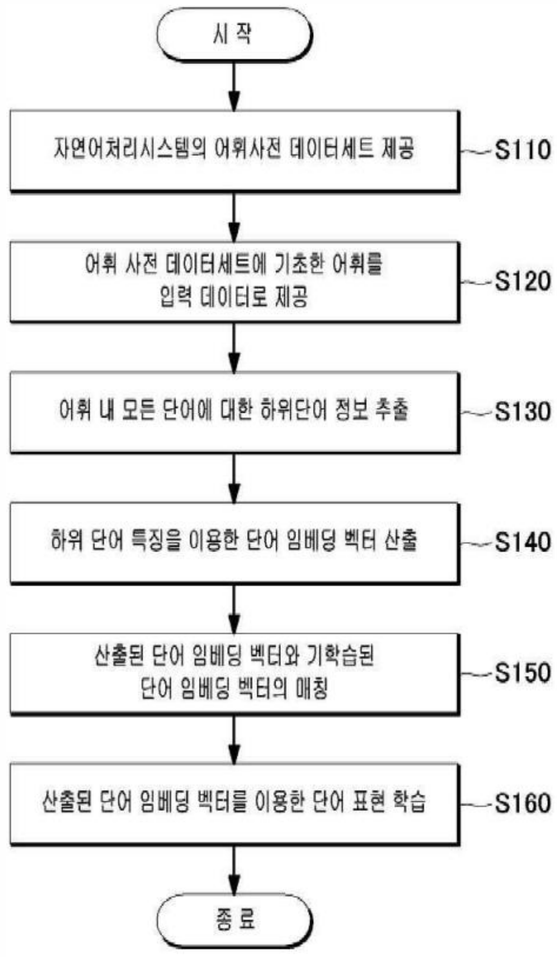


본 기술의 우수성 및 파급 효과

본 기술의 우수성(효과)

- 하위단어 정보를 이용한 단어의 고유한 의미 추론
  - 자연어 처리 시스템은 적어도 하나 이상의 단어를 포함하는 어휘 및 기학습된 단어 임베딩 정보를 포함하는 어휘 사전 데이터세트를 제공함
  - 어휘 사전 데이터세트에 기초한 어휘가 단어 표현 모델에 입력 데이터로 제공되면, 입력 데이터에 존재하는 단어들에 대한 하위 단어 정보를 추출하고, 추출된 하위 단어 정보를 이용하여 하위 단어 특징 벡터들을 생성한 후 이들을 결합하여 단어 임베딩 벡터를 산출함
  - 단어 표현 모델은 단어 임베딩 벡터와 해당 단어의 기학습된 단어 임베딩 정보를 매칭하여, 기학습된 단어 임베딩을 산출된 단어 임베딩 벡터로 대체하여 해당 단어에 대한 단어 표현을 학습함
  - 따라서, 기존에 대형 코퍼스에서 비지도 학습 방식으로 단어 표현을 학습하던 방식과 다르게, 기학습된 단어 임베딩 정보를 포함하는 어휘사전 데이터세트를 이용하여 어휘에 포함된 모든 단어에 대한 단어 표현을 학습함
  - 이후 단어 표현 모델에 미등록 단어가 입력 데이터로 제공되면, 미등록 단어에 대해 추출한 하위 단어 정보를 이용하여 미등록 단어의 단어 임베딩 벡터를 산출하고, 산출된 벡터에 기초한 벡터 연산을 통해 단어 임베딩 벡터 간 유사도를 계산하여 미등록 단어의 이웃 단어를 추출하여 고유 의미를 추론함
  - 본 기술에 따른 단어 표현 모델의 단어 유사성 평가 결과, 하위 단어 정보를 추출하는데 유용하며, 대규모 코퍼스에서 단어 표현을 학습하는 FastText보다 더 우수한 학습 성능을 보여줌

단어 표현 모델의 학습 방법



단어 유사성 평가 결과

Language	Dataset	word2vec	FastText	Mimick	GWR⁻	GWR
Ar	WS353	34.9	**52.1**	44.9	36.3	43.8
De	GUR350	34.6	48.7	47.6	35.4	**53.2**
	ZG222	16.2	35.4	36.5	16.4	**38.7**
En	SL999	29.5	27.7	**30.3**	28.6	29.8
	WS353	63.1	64.1	63.7	63.9	**64.6**
Es	MC30	40.6	**72.1**	66.8	41.4	69.8
	WS353	29.5	50.0	40.7	30.8	**50.2**
Fr	RG65	46.8	58.6	51.3	48.1	**60.5**
Ru	HJ	30.3	**60.1**	42.1	31.4	45.5

적용 제품 및 파급 효과

- 자연어 처리 시스템
- 본 기술을 통해 개방 어휘 환경에서도 미등록 단어가 가지고 있는 고유한 의미를 정확히 추출할 수 있는 자연어 처리 시스템을 제공할 수 있음

지식재산권 현황

발명의 명칭	출원/등록번호	출원/등록일자
자연어 처리 시스템 및 자연어 처리에서의 단어 표현 방법	10-2260646	2021.05.31.
패밀리 특허 현황	**패밀리 국가**	
-	-	

문 의 | 고려대학교 산학협력단 기술사업화센터 권성진 02-3290-5832, giannie77@korea.ac.kr