DL 스터디&프로젝트

# [Euron 중급 세션 18주차] 자연어 처리의 모든 것 1. 자연어 처리의 시작

by 공부하자_ 2024. 1. 8.

## 자연어 처리의 모든 것

## 1. 자연어 처리의 시작

### 📌자연어 처리 개요 - 자연어 처리 활용 분야와 트렌드

**핵심어: Natural language processing(자연어처리), Text mining, Information retrieval, RNN, Transformer, Self-Supervised Learning**

**자연어 처리 활용 분야**

1. 자연어 처리(NLP)

Natural Language Understanding(NLU): 컴퓨터가 주어진 단어나 문장, 보다 더 긴 문단이나 글을 이해하는 과정

Natural Language Generation(NLG): 자연어를 상황에 따라 적절히 생성하는 과정

- low-level parsing

- Tokenization: 어떤 문장을 이해하기 위해 문장을 이루는 각 단어들을 정보 단위로 생각하며 이때 단어를 보통 토큰이라 함. 이때 주어진 문장을 단어 단위로 쪼개나가는 과정을 Tokenization이라 부름
- stemming: 수많은 어미의 변화 속에서 모두 같은 의미를 나타내는 단어라는 것을 컴퓨터가 인식할 수 있어야 하는데, 단어의 다양한 의미변화를 없애고 그 의미만을 보존하는 단어의 어근을 추출하는 것이 stemming
- Tokenization이나 stemming 등은 각 단어를 준비하기 위한 의미 단위로 볼 때 가장 낮은 레벨의 태스크

- Word and phrase level
  - NER: 단일 단어 혹은 여러 단어로 이루어진 고유 명사를 인식하는 태스크
  - POS(Part of speech) tagging: 워드들이 문장 내에서 품사나 성분이 무엇인지 알아내는 태스크(주어인지 목적어인지, 부사구인지 형용사구인지, 형용사라면 어디를 꾸며주는지 등)

- Sentence level
  - 감성분석: 어떤 문장(리뷰 등)을 보고 긍정 어조인지 부정 어조인지 파악
  - 기계번역: 영어를 전체적으로 이해하고 한글 문장으로 번역할 때 각 단어별로 적절한 한글로의 번역과 문법에서의 어순을 잘 고려해야 함

- Multi-sentence and Paragraph level
  - Entailment Prediction: 두 문장 간의 논리적인 내표 혹은 모순관계를 예측(논리적으로 양립가능한지)
  - 독해기반 질의응답: 구글 등에서 질문을 서치했을 때, 질문에 있는 키워드들이 포함된 문서들을 검색한 후 그 문서로부터 독해를 통해 주어진 질문에 대한 정답을 알아내서 사용자에게 직접적으로 제시하는 기술
  - 챗봇: 대화를 수행할 수 있는 자연어처리 기술
  - 요약: 주어진 문서를 한 줄로 자동 요약하는 기술

2. 텍스트 마이닝

- 빅데이터 분석과 관련되는 경우가 많음. 과거 자료에서 특정 키워드의 빈도를 시간에 따라 분석함으로써 트렌드를 도출하거나 소비자의 반응을 얻어냄

2024.01.01

[Euron 중급 세션 16주차]…
2023.12.25

백엔드 프로젝트 12주차 - …
2023.12.23

최근댓글

백엔드 프로젝트 12주차 - …

오늘 하루 고생 많으셨습니다.

좋은 글 잘 보고 가요! 감사…

좋은 글 잘 보고 가요! 감사…

태그

데이터분석, 판다스입문, 딥러닝교과서, 데이터사이언스, 이지스퍼블리싱, 판다스, 딥러닝스터디, Doit, pandas, bda

- 서로 다른 키워드지만 비슷한 의미를 가지는 키워드들을 그루핑해서 분석해야 하고, 자동으로 수행할 수 있는 기법으로써 토픽 모델링이나 문서 군집화같은 기술이 존재. 이러한 기술을 통해 사람들이 어떤 상품의 세부적인 요소를 이야기하고 있고 그에 대한 의견은 어떠한지 빠르게 얻어낼 수 있음
- 빅데이터에 기반한 사회과학과도 밀접한 관련이 있는데, 신조어와 사회현상 파악, 키워드와 현대사람들 패턴 변화 파악 등 사회과학적 인사이트를 발견하는 데 텍스트마이닝이 사용됨

3. 정보 검색

- 구글이나 네이버 등에서 사용되는 검색 기술을 연구, NLP나 텍스트 마이닝보다는 어느 정도 성숙한 상태로 기술 발전이 상대적으로 느린 분야
- 세부 분야로 추천 시스템이 존재, 사용자가 수동으로 검색하는 과정을 자동화해서 보다 적극적이고 자동화된 검색 시스템의 새로운 형태 >> 상업적으로도 큰 임팩트를 가져옴

**NLP의 발전 과정**

- 워드임베딩: 2~3년 전까지만 해도 컴퓨터비전이 훨씬 빠르게 발전하고 있었으나 자연어처리도 꾸준히 발전하고 있는 중이었음. 이때 대부분의 머신러닝, 딥러닝은 숫자 데이터를 다루기 때문에 자연어처리를 적용하기 위해서는 텍스트 데이터를 단어 단위로 분리하고 각 단어를 특정한 차원으로 이루어진 벡터로 표현하는 과정을 거치게 됨. 이는 워드를 벡터공간의 한 점으로 나타낸다는 점에서 워드임베딩이라고 부르게 되었음.
- 시퀀스 데이터를 처리하는 데 특화된 모델 구조로써 RNN이 자연어처리의 핵심모델로 자리잡게 됨 >> RNN에 LSTM, GRU 등의 모델들이 많이 사용됨. 특히 기계번역 부분에서, 룰을 가지고 어순 등을 하나하나 적용하는 것보다 각 원본과 번역본 학습 데이터로 훈련시킨 RNN 모델을 사용하는 것이 훨씬 성능을 높이는 데 기여함
- RNN으로 이미 성능이 높여진 상태에서 이를 더 높일 수 있는 transformer 모델이 발표됨.  현재 대부분의 자연어처리 모델은 transformer 모델을 기반으로 하며, 자연어처리 뿐만이 아니라 영상처리, 시계열 예측 등 다양한 분야에도 활발히 적용되어 성능향상을 이루어내고 있음
- transformer 이전에는 서로 다른 자연어 처리 태스크 별로 태스크에 특화된 딥러닝 모델이 각기 따로 존재해옴. 하지만 요즘 transformer 모델에서 사용된 핵심 모듈인 self attention 모듈을 쌓아나가며 모델 크기를 키우고, 큰 구조의 변화 없이 원하는 태스크의 전이 학습의 형태로 적용했을 때 전이 학습의 형태

로 적용했을 때 특화된 모델들보다 월등히 뛰어난 형태로 발전함.

- 최근 들어서는 자가지도학습이 유행하고 있음. 자가지도학습은 입력 문장이 주어졌을 때 한 단어를 가리고 앞뒤 문맥을 보고 그걸 맞추게하는 태스크에 해당. 간단하게 설계된 태스크에서도 언어의 문법적인, 의미론적인 지식을 딥러닝이 학습할 수 있게 됨. 이런 자가지도학습의 예시로 BERT, GPT2, 3 등이 존재함. 이러한 모델들은 특정한 태스크만 수행하는 제한적인 인공지능 기술에서 나아가 범용 인공지능 기술로써 현대의 기술이 한 단계 더 발전한 것으로 볼 수 있음. 그런데 이런거 훈련시키려면 돈이 너무 많이 소비되어 막강한 자본이 뒷받침되는 소수 기관에서 이루어짐.

## 📌자연어 처리 개요 - 기존의 자연어 처리 기법

**핵심어: Bag-of-Words, One-hot vector, Naive Bayes Classifier, 문장 분류**

### Bag-of-Words

- NLP나 텍스트 마이닝 분야에서 딥러닝 기술이 적용되기 이전에 많이 활용되던, 단어나 문서를 숫자 형태로 나타내는 가장 간단한 기법이 Bag-of-Words, 그리고 이를 활용한 대표적인 문서 분류 기법이 Naive Bayes Classifier임.
- Vocabulary에 있는 단어들을 범주형 변수로 두고, 이 범주형 변수를 One hot 벡터로 표현. 이때 8차원의 좌표공간에서 어떤 단어의 순서쌍이던지 그 사이 거리는 루트2, 모든 단어 순서쌍의 코사인 유사도는 0. 따라서 단어의 의미에 상관없이 모두가 동일한 관계를 가지는 형태로 벡터를 설정
- 다수의 문장으로 구성된 문서도 One hot 벡터를 확장해서 나타냄(모두 더함으로써) >> Bag-of-Words 벡터
- Vocabulary상에서 존재하는 각 워드별로 어떤 가방을 준비하고, 특정 문장에서 나타난 워드들을 순차적으로 해당하는 가방에 넣어준 후 최종적으로 각 차원에 해당하는 가방에 들어간 워드의 수를 세서 최종 벡터를 나타냄

### Naive Bayes Classifier

- Bag-of-Words 벡터로 나타낸 문서를 정해진 카테고리 혹은 클래스 중에 하나로 분류할 수 있는 대표적 방법

- 문서가 분류될 수 있는 카테고리 혹은 클래스가 C개 있으면, 특정한 문서 d가 C개의 클래스에 속할 확률분포는 P(c|d). 이러한 조건부 확률 분포는 가장 높은 확률을 가지는 클래스 C를 택하는 방식을 통해서 문서 분류를 수행할 수 있게됨.
- P(d|c)는 특정 카테고리 c가 고정되었을 때 문서 d가 나타날 확률, 문서d는 w,부터 마지막 워드 wn까지 동시에 나타나는 동시 사건으로 볼 수 있음. 각 단어가 등장할 확률이, c가 고정되어있는 경우 서로 독립이라고 가정할 수 있다면 각 단어가 나타낼 수 있는 확률을 모두 곱한 형태로 나타낼 수 있음.
- 베이즈 정리에 의해, 최종 계산 값 P(c|d)를 알려면 P(c)값과 각각의 P(w|c)값을 알아야 함. 그리고 가장 큰 P(c|d)를 지니는 클래스가 최종 선택됨
- P(c) 즉 어떤 문서가 주어지기 이전에 각 클래스가 나타날 확률과, 특정 클래스가 고정되어있을 때 각 워드가 나타날 확률을 추정함으로써 Naive Bayes Classifier에서 필요한 파라미터들을 모두 추정할 수 있게 됨

# 📌자연어 처리와 벡터 - Word Embedding (1)Word2Vec

**핵심어: Word Embedding, Word2Vec, window size, Embedding**

### Word Embedding

- 각각의 단어를 특정한 차원의 벡터로 표현할 수 있는 기법
- 자연어가 정보의 기본 단위가 되는 단어의 시퀀스라고 할 때, 어떤 특정한 차원으로 이루어진 공간 상의 한 점, 혹은 그 점의 좌표를 나타내는 벡터로 변환해주는 기법
- 예를 들어 3차원 공간 상에서 각 단어를 한 점으로 나타낸다고 하면, (1, 3, 4)를 cat으로 할당함
- 워드임베딩 그 자체가 머신러닝, 딥러닝 기술 자체로써 학습데이터를 주고 좌표 공간의 차원 수를 미리 정의해주면, 그 이후에 워드임데딩 학습 완료되면 해당 좌표공간 상에서 학습 데이터에서 나타나는 최적의 좌표값, 벡터 표현형을 출력으로 보여주게 됨
- 기본 아이디어는 '비슷한 의미를 가지는 단어가 좌표공간 상에 비슷한 위치 점으로 매핑되도록 함으로써 의미 상의 유사도를 잘 반영한 벡터 표현을 다양한 자연어처리 알고리즘에게 제공해주는 역할을 하게 됨. (캣과 키티는 좌표공간 상에서 가까운 위치에, 햄버거는 두 단어보다 먼 위치에 좌표값을 부여받게 됨)

- 보다 쉽게 자연어처리 태스크 성능을 올리는 여건을 만들어줌 (love, like 비슷한 벡터 표현형 >> 긍정 어조 감정분석 쉬워짐)

**Word2Vec**

- 워드임베딩을 학습하는 방법 중 가장 유명한 방법 중 하나
- 비슷한 의미를 가지는 단어가 좌표공간 상에서 가까운 위치에 매핑되도록 하기 위해, 같은 문장에서 나타난 인접한 단어들 간에 의미가 비슷할 것이라는 가정을 사용함
- 어떤 한 단어가 주변에 등장하는 단어들을 통해 그 의미를 알 수 있다는 사실에서 착안, 주어진 학습 데이터를 바탕으로 특정 단어 주변에 나타나는 단어들의 확률 분포를 예측함.
- 먼저 주어진 문장을 워드별로 분리하는 Tokenization 과정을 수행, 그리고 유니크한 단어들만을 모아서 사전을 구축하게 됨. 사전의 각 단어는 vocabulary 사이즈만큼의 차원을 가지는 one hot 벡터로 나타나짐. 이후 Sliding window 기법을 적용하여, 어떤 한 단어를 중심으로 앞뒤로 나타난 워드 각각과 입출력 단어쌍을 구상하게 됨. 이렇게 주어진 학습데이터에 대해 각 문장 별로 Sliding window를 적용하고 거기서 어떤 중심 단어와 주변 단어 각각을 단어 쌍으로 구성함으로써 Word2Vec의 학습데이터를 구성할 수 있게됨
- 많은 입출력 단어쌍들에 대해 예측 태스크를 수행하는 2개 레이어의 신경망을 만들며, 각 단어가 vocabulary 사이즈만큼의 one hot 벡터로 나타나기 때문에 입력과 출력 레이어의 노드 수는 차원 수가 되고, 입출력이 각각 입력 및 출력 단어에 해당, 가운데 있는 히든레이어의 노드 수는 사용자가 정하는 하이퍼파라미터로써 워드임베딩을 수행하는 좌표공간의 차원 수와 동일한 값으로 설정함.
- 입력 단어의 워드임베딩과 출력 단어의 워드 임베딩을 조정해가며 학습을 진행해가는 것이 Word2Vec의 핵심
- 유사한 벡터 표현형을 가지면 >> 비슷한 의미로 매핑될 가능성이 높아짐

**Word2Vec 성능**

- 어떤 공간에서 두 벡터 포인트 사이의 관계는 두 단어 사이의 관계를 나타냄
    - Woman - Man, Aunt - Uncle, Queen - King 등
- 동일한 관계는 동일한 벡터로 나타남
- Word intrusion detection 태스크: 여러 단어들이 주어져있을 때, 나머지 단어와 그 의미가 가장 상이한 단어 하나를 찾아내는 태

스크

**Word2Vec 응용**

- 자연어를 워드 단위의 벡터로 나타냄
- 기계 번역: 서로 다른 언어 간의 같은 의미를 가지는 워드들의 임베딩 벡터가 쉽게 정합될 수 있도록 하여 번역 성능 높임
- PoS 태깅
- 감정 분석: 각 단어의 긍/부정을 보다 용이하게 파악하도록
- Image Captioning: 주어진 이미지의 상황을 잘 이해하고, 이에 대한 설명을 자연어 형태로 생성

# 📌자연어 처리와 벡터 - Word Embedding (2)GloVe

**핵심어: Word Embedding, Word Representation, GloVe(Global Vectors for Word Representation)**

**GloVe**

- Word2Vec과 더불어 많이 쓰이는 또다른 워드임베딩 방법
- Word2Vec과의 가장 큰 차이점은 각 입력 및 출력 단어 쌍들에 대해, 학습데이터에서 두 단어가 한 윈도우 내에서 총 몇 번 동시에 등장했는지를 사전에 미리 계산하고 새로운 형태의 손실함수를 사용
  - Word2Vec: 특정한 입출력 쌍이 자주 등장한 경우, 그러한 데이터 아이템이 자연스럽게 여러 번에 걸쳐 학습됨으로써 두 워드임베딩 벡터 간의 내적값이 이에 비례하여(학습이 빈번할수록) 커지도록 하는 학습 방식
  - Glove: 애초에 어떤 단어 쌍이 동시에 등장한 횟수를 미리 계산하고, 이에 대한 로그값을 취한 값을 직접적인 두 단어 간의 내적 값에 계산 및 학습 진행 >> 중복되는 계산 줄여줌, Word2Vec보다 빠르게 수행
- GloVe 모델: 단어들 사이의 문법적 의미와 관계들까지 비슷하면 벡터가 유사함

공감