자연어 처리의 이해

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

01

APPLICATION

ARTIFICIAL INTELLIGENCE

Notice

THINKING

생각해보기

✓ 자연어 데이터 전처리란 무엇일까요?

학습목표

GOALS

1 자연어란 무엇인지 설명할 수 있다.

2 자연어 처리란 무엇인지 설명할 수 있다.

3 한국어의 특성에 대해 이해하고 설명할 수 있다.

4 한국어에서의 자연어 처리 시 어려움에 대해 파악하고 해결할 수 있다.

5 전통적인 자연어 처리에 대해 이해하고 설명할 수 있다.

6 자연어 처리의 발전에 대해 이해하고 설명할 수 있다.

1 자연어 처리 소개

2 영어 자연어 처리 개요

3 한국어 자연어 처리 개요

CONTENTS

학습내용

APPLICATION

01 자연어 처리 소개

01 자연어란?

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

자연어 <-> 프로그래밍언어

- 우리가 일상적으로 듣고 쓰고 말하고 읽는 모든 언어
- 인간으로써 우리는 언어를 통해 생각이나 감정을 표현
- 파이썬, C++ 등 인공 언어와 반대되는 개념

자연어 처리

NLP (Natural Language Processing)

인공지능의 한 분야로, 사람의 언어 현상을 컴퓨터와 같은 기계를 사용해 다루는 작업

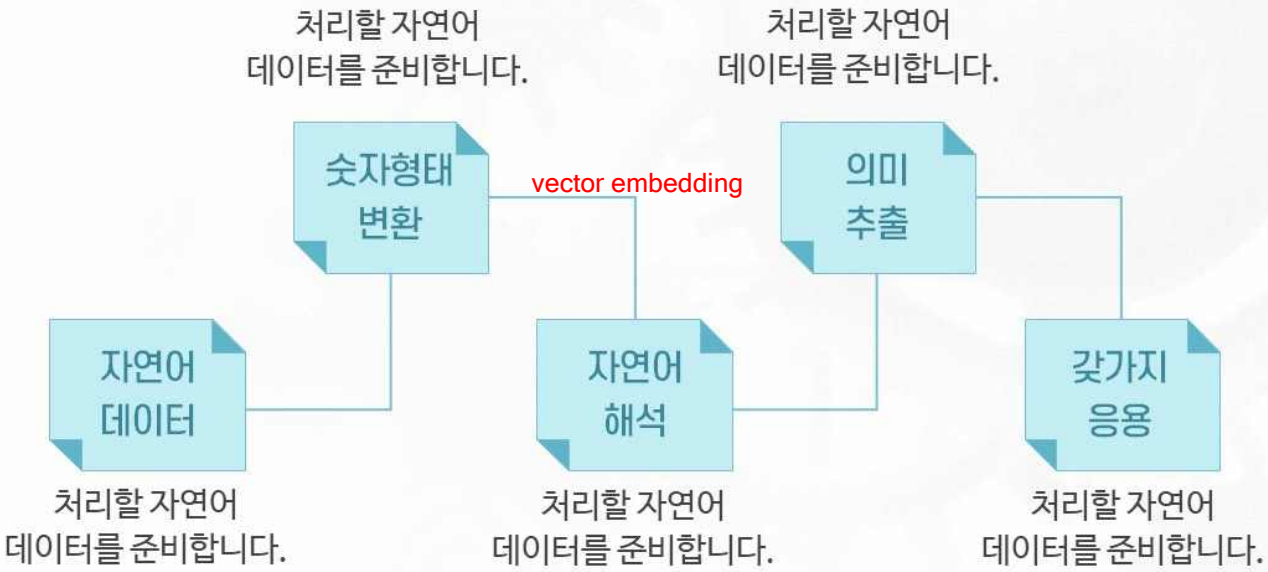
02 자연어 처리란?

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

자연어 처리의 최종목표는

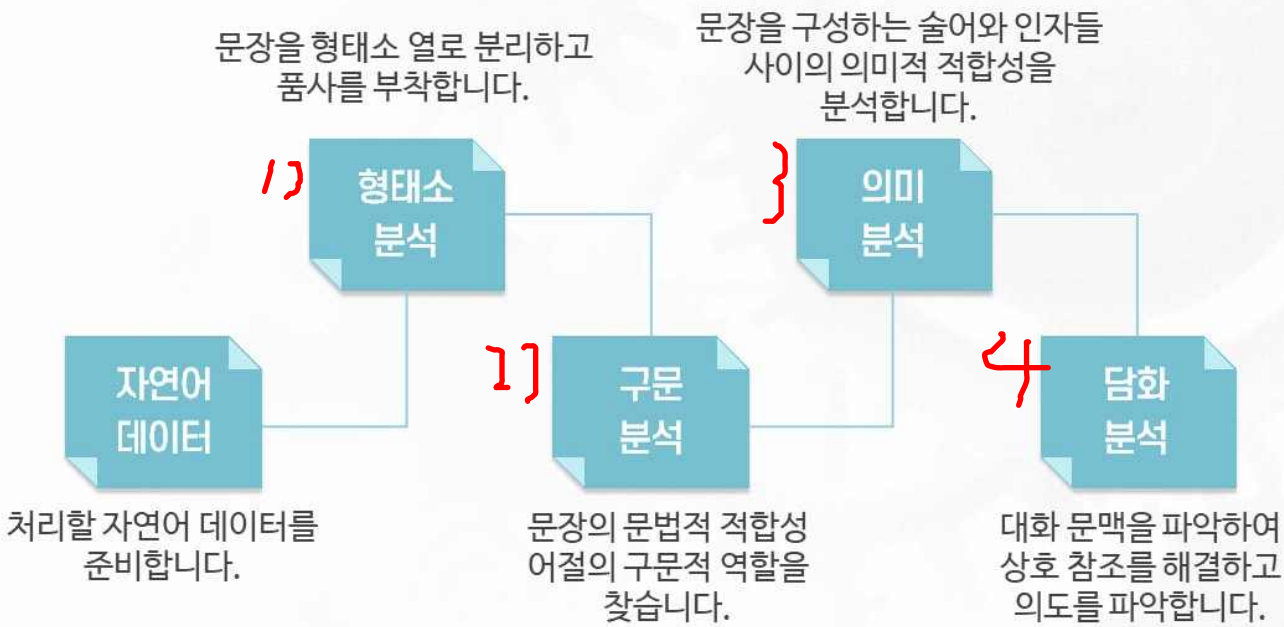
> "컴퓨터가 사람의 언어를 이해하고
> 여러가지 문제를 해결할 수 있도록 하는 것"

03 일반적인 자연어 처리 과정



음성 텍스트 영상

04 자연어 분석 단계



05 자연어 처리 사용 분야

텍스트 분류

챗봇

감정 분석

특정언어에서 다른 언어로의 번역

NMT

텍스트 요약

질문 응답 시스템

출처: 퍼블릭에이아이(www.public.co.kr)

06 자연어 처리의 어려움

APPLICATION

02 한국어 자연어 처리 개요

01 한국어 문법 용어

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

음절

한글 자모 첫소리와 가운뎃소리 글자, 또는
첫소리, 가운뎃소리, 끝소리 글자로 이루어진 한글의 단위

- 한국어는 초성, 중성, 종성을 조합해서 총 11172개의 글자를 만들 수 있음

어근

실질적인 의미를 나타내는 말의
중심이 되는 부분

ex) '덮개'의 덮, '어른스럽다'의 어른

어간

뜻을 가진 가장 작은 말의 단위

ex) '보니', '보다', '보고'에서 '보-' 부분

01 한국어 문법 용어

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

접사

독립적으로 쓰이지 못하고 어근이나 어간 등에 붙어
의미나 품사를 바꿈

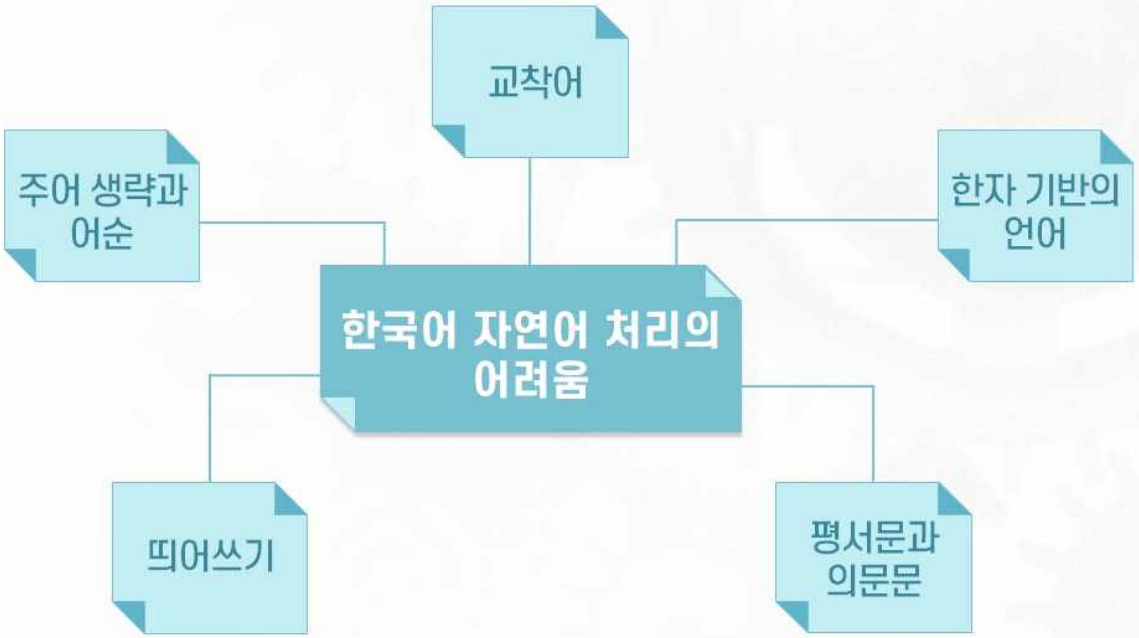
ex) '맨'몸, 지우'개', 먹'이'

형태소

뜻을 가진 가장 작은 말의 단위

ex) 아기가 밥을 먹는다 → 아기/가/ 밥/을/ 먹/는/다

02 한국어 자연어 처리의 어려움



03 교착어란?

교착어

언어의 유형론적 분류 중 하나로,

단어의 중심이 되는 **형태소(어근)**에

접사를 비롯한 **다른 형태소들이 덧붙어 단어가 구성**되는 것이 특징

03 교착어

하나의 어근에 의해 다양한 단어들이 생성됨

ex) '풀' 이라는 어근에 의해 〈먹다, 먹었다, 먹히다, 먹히었다, 먹었겠더라〉 등등 다양한 단어들이 생성

접사에 따라 단어의 역할이 정의됨

ex) '다리가 아파서 다리를 주물렀더니 다리의 통증이 괜찮아졌다'라는 문장에서 컴퓨터는 〈'다리가', '다리를', '다리의'〉 세 단어를 모두 다른 단어로 간주할 가능성이 있음

04 평서문과 의문문

평서문

- 화자가 문장의 내용을 객관적으로 진술하는 문장
- 일반적인 서술형 문장

&

의문문

- 화자가 청자에게 질문을 하여 답을 요구하는 문장

"어디 갔다 왔니? vs 어디 갔다 왔어? Vs 어디 갔다 왔어."

띄어쓰기 에 대한 **표준**은 계속 바뀌어 왔고

띄어쓰기 적용 방식은 매우 까다로운 편임

05 띄어쓰기

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

띄어쓰기 를 하지 않아도 의미가 전달되는 경우도 많고,

실제로 잘 사용하지 않는 경우도 많음

ex) 이번휴가때 어디로놀러가?

주어 생략

한국어에서는 주어를 생략하는 경우가 많은데,
컴퓨터는 사람과 달리 생략된 정보를 매꿀 수 없어
문장의 정확한 의미 파악이 힘듦

ex) (너) 밥 먹어

어순

접사에 따라 단어의 역할이 정의 되는
한국어에서는 어순이 중요하지 않음

ex) (너) 밥 먹었어? (너)

06 주어 생략과 어순

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

나는 기쁘게 결승선을 통과했다.

나는 결승선을 기쁘게 통과했다.

통과했다 결승선을 나는 기쁘게.

기쁘게 결승선을 통과했다 나는.

결승선을 기쁘게 통과했다 나는.

표어 문자와 표의 문자

한자 & 한글

- 하나의 문자로 단어 또는 형태소를 나타내는 표어 문자
- 읽는 방법은 같지만 문자의 모양이나 의미는 다른 경우가 있음
- 사람의 말소리를 기호로 나타낸 표음 문자

- 표어 문자인 한자를 표음 문자인 한글로 표현하는 과정에서 정보의 손실이 존재함
- 같은 글자처럼 보이지만 하나의 음절이 다른 의미를 지니는 경우가 존재함

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

1 각종 분야에서의 딥러닝 언어 모델 구축

특허분야에서의 유사 특허 분류 업무를 진행하는
딥러닝 언어 모델

2 각종 분야에서의 딥러닝 언어 모델 구축

기존 일방향 정보 소통의 AI스피커가 아닌 자연어 처리 기술을
도입해 양방향 소통 및 상호작용이 가능해짐

03 자연어 처리의 발전 과정

01 전통적인 자연어 처리 기술

알고리즘 기반 후보 생성

전통적인 자연어 처리 기술들은 알고리즘을 이용해
여러 개의 후보를 생성하고 확률적인 방법으로
모호함을 해소함

알고리즘 기반 후보 생성

형태소 분석

- 태뷸러 파싱 알고리즘
- 어절로 부터 모든 가능한 형태소 후보를 생성

의존 구조 분석

- 어절과 어절 사이의 관계를 의존 문법으로 기술

01 전통적인 자연어 처리 기술

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

확률 기반 애매성 해소

대용량의 통계데이터를 바탕으로 확률적 선택을 통해 모호함을 해소하는 과정이 필요함

형태소 분석의 경우 애매성을 해소하는 전통적인 방법으로 HMM(Hidden Markov Model)이 존재함

1 머신러닝의 도입

- 자연어 처리 분야에서도 딥러닝 모델을 통한 자연어 처리 연구가 활발히 진행되고 있음
- 2013년 word2vec의 발전 이후 딥러닝 연구자들의 자연어 처리에 대한 이해도가 높아짐

02 자연어 처리 기술의 발전

DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

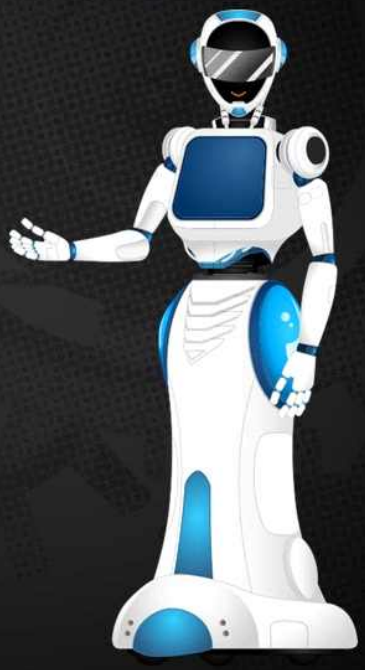
2 딥러닝을 통한 자연어 처리

- 합성곱 신경망(Convolutional Neural Network, CNN)
- 순환 신경망 (Recurrent Neural Network, RNN)
- 강화 학습 (Reinforcement Learning)

SUMMARY

학습정리

- 자연어와 자연어 처리
- 자연어 처리의 어려움
- 한국어에서의 자연어 처리와 어려움
- 전통적인 자연어 처리 방법
- 딥러닝과 자연어 처리 기술의 발전



DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

EXPANSION

확장하기

1. 자연어 처리는 실제 생활의 어떤 분야로 **활용**이 가능할까요?
2. 자연어 처리에서 **고려해야 할 문제**로는 어떤 것들이 있을까요?
3. 한국어에서의 **자연어 처리가 더욱 어려운 이유**는 무엇일까요?
4. **형태소 분석 과정에서 모호함을 처리하기 위한 방법**으로 어떤 것이 있을까요?

참고 문헌

REFERENCE

◆ 참고 논문
- Young, T, Hazarika, D., Poria, S., & Cambria, E. (2017). Recent Trends in Deep Learning Based Natural Language Processing. arXov preprint arXiv:1708.02709

◆ 참고 사이트
- 용어들에 대한 정의: https://ko.wikipedia.org/wiki.
- 퍼블릭에이아이(www.public.co.kr)

◆ 참고 서적
- 김기현, 『김기현의 자연어 처리 딥러닝 캠프 파이토치 편』, 한빛미디어, 2019
- 잘라지 트하나키, 『파이썬 자연어 처리의 이론과 실제』, 에이콘, 2017

서체 출처 : 에스코어드림체-㈜에스코어, 나눔글꼴체-㈜네이버