ch1
자연어처리 : 자연어를 받아 컴퓨터에서 이해하고 다시 사용자에게 이해 가능한 언어를 생성해 내는 일련의 과정 ==> 자연어를 입, 출력으로 사용하는 컴퓨터에 사용되는 처리 과정.

자연어 입력 (NLU) : 문자로 된 언어를 이력으로 직접 받아들여, 목적에 맞게 내부적으로 처리해내는 과정. (예: 자연어를 입력으로 하여 다른 프로그램을 호출하는 시스템(빅스비, 시리))
자연어 생성 (NLG) : 주어진 정보를 바탕으로 문장을 생성하여 사용자에게 자연어로 응답을 돌려준다. (예: 날씨 앱에서 자연어로 날씨를 설명)

응용 분야 : 전산언어학 - 주로 언어의 규칙 등을 찾기 위해 규칙 기반이나 통계 기반의 언어를 많이 진행, 최근 딥러닝 도입. 기계번역, 기사 요약 서비스, 음성인식(특히, 동음이의어를 처리할 때 자연어처리 사용. ('무리'라는 발음에서 '물이'와 '무리' 중 앞, 뒤 문맥에 맞춰 뜻을 선택하는 것.)), 개인 비서 서비스

왜 어렵지?
처리해야할 데이터가 수치화된 값이 아닌 '자연어'이기 때문 ==> 해석이 누구나 같음, 하지만 자연어는 중의성, 규칙의 예외, 언어의유연성과 확장성 때문에 힘듬.

언어의 중의성-동음이의어나 반어법등 맥락에 따라 해석의 여지가 달라지는 것 때문에 해석에 어려움을 줌.

규칙의 예외-한글만 봐도 언어 문법에 예외가 존재, 또한, 숙어나 속담 같은 경우는 뜻이 아예 달라져 해석에 어려움, 문법에 맞게 구성되어있지만. 뜻이 없는 문장도 어려움에 속함.

언어의 유연성과 확장성-단어와 소리를 조합하여 만들 수 있는 문자의 수와 길이가 무한함. 언어를 모델화하여 처리하는 데 있어 이 단어가 어느 구에 속하는지 등 해석하는데 어려움이 있음. 신조어나 고어등 언어의 뜻이 바뀌거나 새로운 단어가 만들어지고, 줄임말과 같이 새로게 계속 생기기 때문에 대응하기가 어렵다.

자연어처리 연구의 패러다임
규칙 기반
언어의 문법적인 규칙을 사전에 정의해두고 그것에 기반하여 자연어를 처리하는 방식. 가장 전통적인 방식. 1954년에 조지타운 대학과 IBM이 공동으로 개발한 러시아어-영어 번역기가 시초.
기계번역을 규칙 기반으로 처리하는 과정 : 핵심이 되는 단어들을 사전으로 번역한 다음, 원문장에서 문법적인 규칙을 찾아낸 후 대응하는 번역한 언어의 규칙을 불러와 단어와 단어 사이를 이어 준다. 예시) "Send a message to Susan that I will be lae for meeting."을 입력으로 받는다면 Send, Message, late, meeting이 핵심 단어이고, 명령어, that 등의 문법을 찾아 우리말로 대응하여 해석하는 것이다.
단점 : 규칙을 사전에 직접 구축해야 한다. ==>대응할 수 있는 문장의 종류가 제한되거나 정확도가 매우 떨어진다.

하지만, 문접적인 부분을 처리할때 다른 방법과 융합해서 사용

통계 기반
규칙 기반방법의 한계를 극복하기 위해 제시된 방법.
‘조건부 확률’을 중심으로 사용. ex) 문장을 단어(혹은 형태소)별로 나눈다. ==> 문장을 완성시켜 나갈 때, 앞, 뒤에 등장한 단어라는 이미 일어난 사건에 대해 다음에 어떤 단어가 나올 확률이 가장 높은지를 계산하여 구한다. 가장 확률이 높은 단어가 가장 자연스러운 문장이 될 확룰이 높아진다.
단어 사이의 상관관계를 이용한 것

딥러닝 기반
-직접적인 알고리즘을 개발하는 것이불가능 할 때 문제 해결을 위한 프로그램을 개발.
사용되는 가중치를 계속해서 갱신시키는 것
여기서 딥러닝은 기계학습 중, 신경망 구조에서 뉴런의 층 수를 여러단계로 만든 것
연산할 가중치의 개수가 너무 많아 연산의 흐름을 디자인한 프로그래머조차 정확하게 어떤 가중치가 무엇을 의미하는지 알 수 없다. ==> 어떻게 이런 결과를 냈는지 알 수 없는 블랙박스가 되었다.

딥러닝을 사용하는 자연어 처리 연구
1장에서는 간단한 흐름 설명
1. 자연어처리를 도입하는 목적을 결정
2. 해당 목적과 관련한 학습 데이터(코퍼스)를 구축.
3. 학습 데이터를 통해 학습시킬 모델 작성
4. 모델을 코퍼스로 학습시킨다.
5. 모델 검증하고 피드백
6. 실전 투입

-단어 임베딩(word embedding)
자연어로 되어 있는 문장을 컴퓨터가 받아들일 수 있도록 하는 문장의 전처리 과정 중 하나이다. 특히, 단어(형태소) 단위로 문장을 분해할 때 주로 사용.
각 단어(형태소)를 벡터로 변환하여 비슷한 단어들은 거리가 가깝게, 비슷한 관계가 있는 단어쌍은 거리와 방향을 비슷하게 한다.(예시: man-woman, king-queen은 남,여일때 지칭하는 단어라는 비슷한 관계가 있어 거리와 방향을 비슷하게 함.)

-코퍼스
우리말로 말뭉치를 의미. 매우 많은 수의 문장의 모음.
모델을 학습 시킬 때 입력 문장과 쌍으로 결과를 넣어줘야 하므로 목적에 따라 변형시킨다.
ex) 문자 메시지 전송을 목적으로 하면, 메시지의 내용중 그대로 전달할 부분에 대한 정보를 기입해 놓는다.
코퍼스에 따라 모델의 성능이 좌우되므로 매우 중요함
모델의 성능을 올리기 위해 다시 새로운 기준으로 코퍼스를 필터링 하는 것도 자연어처리 관

련 연구의 한 분야.

-모델
모델을 구축한다는 것은 어떤 학습 과정을 거치게 하는지에 대한 고민
-어떤 단어를 처리할 때 앞, 뒤 단어들을 중점적으로 봐야할 것 같으면, 모델이 분석할 때 단어와 그 중심을 중점적으로 보도록 구조를 작성. 전체적인 맥락을 봐야한다면, 전체를 아우르는 데이터 간 연결을 만들어준다. ==> 이처럼 목적에 따라 다양한 학습 과정을 거치게 할 수 있다.

과정 뿐만 아니라 출력 형식을 지정하는 것도 중요
-문장의 긍,부정 여부이면 0과 1로 출력하는 모델 작성
-형태소를 분석하는 모델이면 각 형태소에 라벨을 붙이게 한다.