자연어처리 모델 응용 분야

1. 언어 모델 응용 분야
    1) 기계 번역(Machine Translation)
    2) 오타 교정(Spell Correction)
    3) 음성 인식(Speech Recognition)
    4) 검색어 추천(Keyword Search Recommendation)
2. 자연어처리 대표적인 응용 분야
    1) 의미론적 유사도 측정
    2) 기계 독해
    3) 관계 추출
    4) 개체명 인식
    5) 감성 분석
    6) 주제 분류

1. 언어 모델 응용 분야

- 언어 모델은 **단어 시퀀스에 확률을 할당**하는 일을 하는 모델, 확률을 통해 가장 자연스러운 단어 시퀀스를 찾아낸 것
- 단어 시퀀스에 확률을 할당하게 하기 위해서 가장 보편적으로 사용되는 방법은 언어 모델이 **이전 단어들이 주어졌을 때 다음 단어를 예측**하는 방법.

1) 기계 번역(Machine Translation)

아래와 같이 번역기를 돌린 2가지 문장이 있다고 가정했을 때, 좌측 문장이 우측 문장보다는 자연스러운 문장임

→ 언어 모델은 좌측 문장이 우측 문장보다 높은 확률을 갖는다고 판단함

P("나는 지하철을 탔다") > P("나는 지하철을 태웠다")

2) 오타 교정(Spell Correction)

2가지 예시 문장을 보면 오타가 발생한 우측 문장보다는 좌측 문장이 자연스러운 문장임
→ 언어 모델은 좌측 문장의 확률이 우측 문장보다 높다고 판단함

P("자연어처리 방법론을 알아갔다") > P("자연어처리 방법론을 랄아갔다")

3) 음성 인식(Speech Recognition)

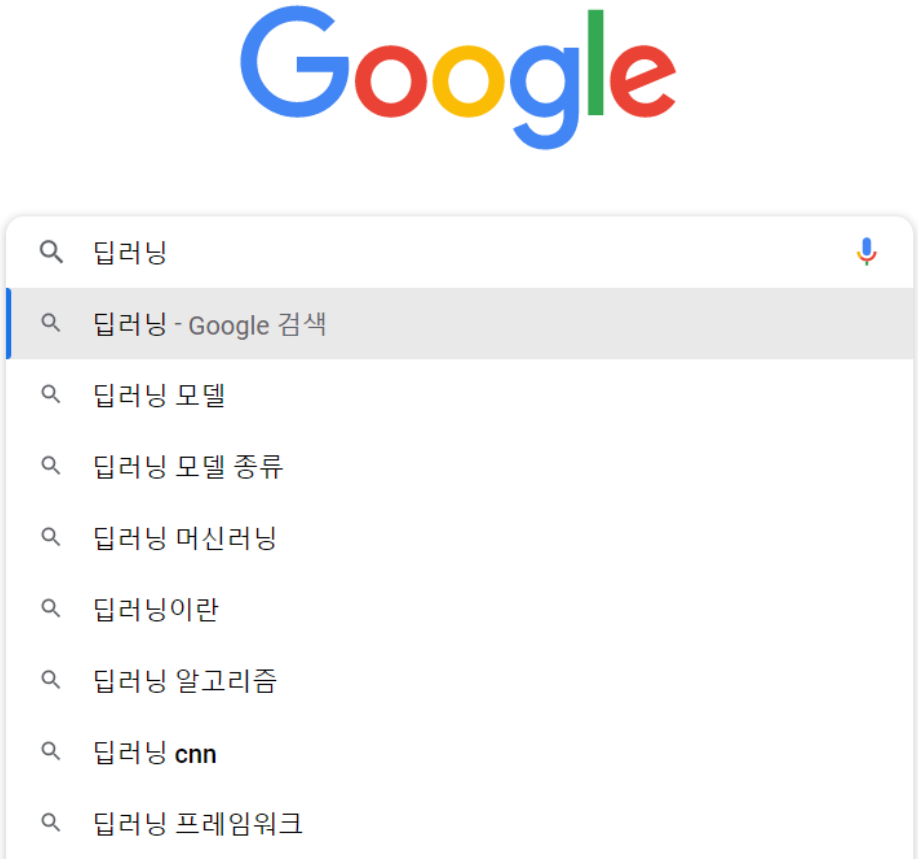
"신라"는 소리 내어 발음하면 [실라]라고 발음하게 됨.

→ 언어 모델은 실제 텍스트와 발음 간의 차이를 교정해 주는 데 활용할 수 있음.

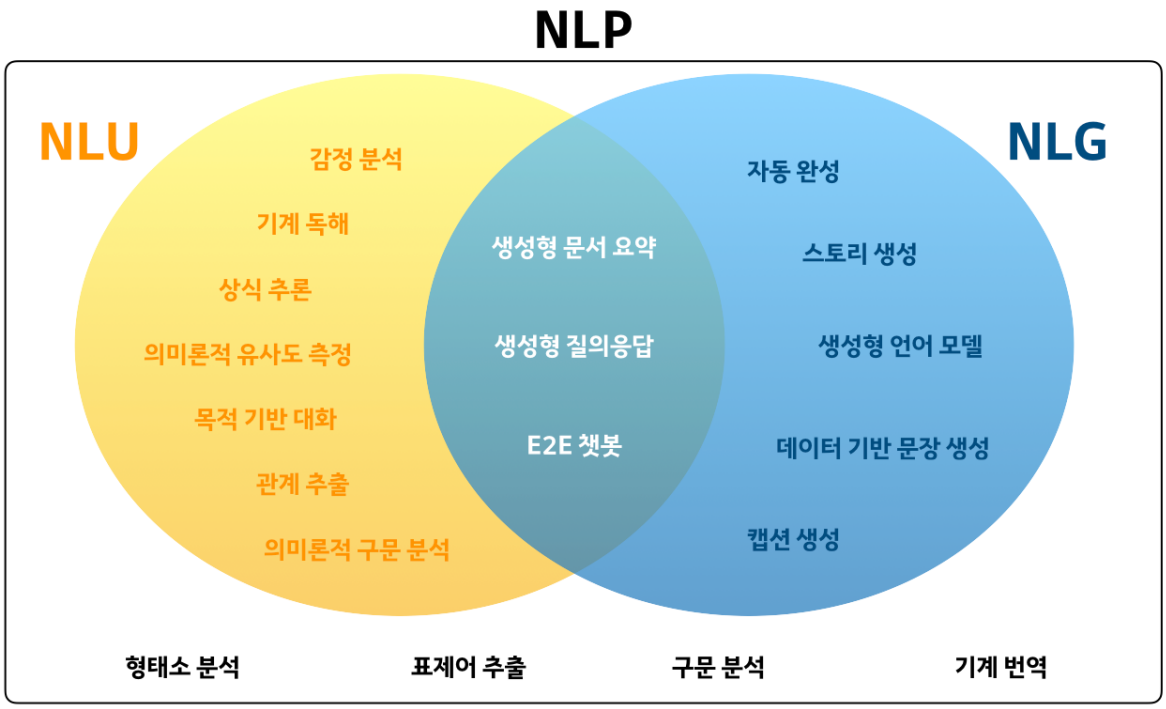
P("신라의 달밤") > P("실라의 달밤")

4) 검색어 추천(Keyword Search Recommendation)

언어 모델은 주어진 텍스트 다음의 단어 시퀀스를 예상하기 때문에 검색어 입력시 다음으로 올만한 자연스러운 검색어들을 추천할 수 있음



2. 자연어처리 대표적인 응용 분야



- NLP(Natural Language Processing) 자연어 처리
- NLU(Natural Language Understanding) 자연어 이해 : 자연어 형태의 문장을 이해하는 기술
  - 사람-기계 상호작용이 필수인 경우 NLU는 핵심 기술
  - ex) 구글에서 NLU 기술을 접목해 기존 키워드 매칭 방식과 비교해 더 나은 검색 서비스를 제공함.
- NLG(Natural Language Generation) 자연어 생성 : 자연어 문장을 생성하는 기술

1) 의미론적 유사도 측정

- **STS**(Semantic Textual Similarity)
  - **텍스트의 의미적 유사도**를 측정하는 문제
  - 모델이 의미상 두 문장의 유사도를 얼마나 잘 잡아내는지 또는 문장의 의미적 표현을 얼마나 잘 구현하는지 평가하는데 일반적으로 사용됨
  - **다양한 어플리케이션에 활용**될 수 있음
  - ex) 질문 사이트 : 여러 질문들을 검색할 때, 사실상 의미적으로 중복된 질의임을 인식할 수 있다면 검색 처리 과정의 많은 비용이 절감됨

    - 자연어 이해(NLU)에서 STS는 핵심 과제이며, 많은 NLP 응용 프로그램 및 관련 영역의 기본 작업 됨.
- 데이터셋 예시
    - KorSTR

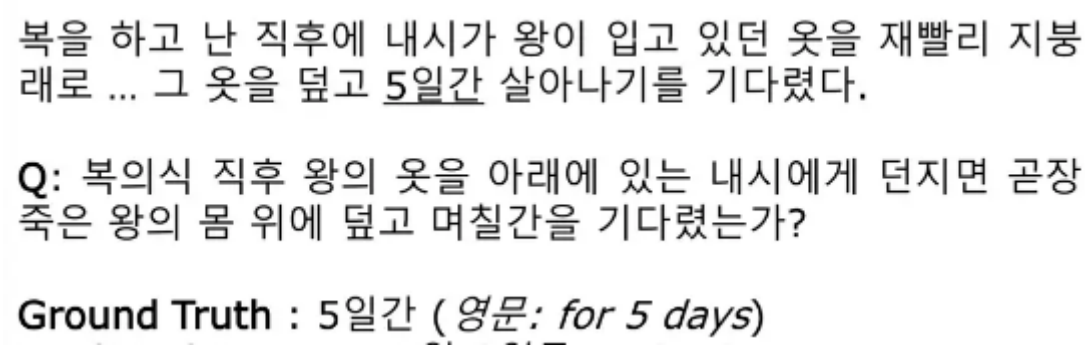
Example	English Translation	Label
한 남자가 음식을 먹고 있다. 한 남자가 뭔가를 먹고 있다.	A man is eating food. A man is eating something.	4.2
한 비행기가 착륙하고 있다. 애니메이션화된 비행기 하나가 착륙하고 있다.	A plane is landing. A animated airplane is landing.	2.8
한 여성이 고기를 요리하고 있다. 한 남자가 말하고 있다.	A woman is cooking meat. A man is speaking.	0.0

2) 기계 독해

- MRC(Machine Reading Comprehension)
    - 자연어로 주어진 질문과 주어진 대상 문서의 내용을 기계가 이해하고 대답하는 문제
    - 모델은 Context, Query, Answer에 대해서 학습을 하고, 실제 사용 단계에서는 Qeury가 주어졌을 때 Context 내에서 Query에 대한 Answer를 찾음

      예시
    - 인공지능 스피커, 검색엔진, 챗봇 등에서 이러한 방식을 사용함
- 데이터셋 예시

  Korquad



- 이전에는 텍스트로만 주어진 context에서 answer을 찾는 태스크가 주를 이뤘으나 이제는 표나 리스트 등으로 구조화된 문서에서도 필요한 정보를 찾아낼 수 있어야 함
  → **TableQA 같은 표 내부에서 정답을 찾는** 태스크 및 모델들에 대한 연구가 이어짐

3) 관계 추출

- RE(Relation Extraction)
  - 텍스트에서 단어들 간의 의미론적 관계를 식별하는 것
  - 관계 추출은 지식 그래프 구축을 위한 핵심 구성 요소로, 구조화된 검색, 감정 분석, 질문 답변하기, 요약과 같은 자연어처리 응용 프로그램에서 중요함
- 데이터셋 예시

  KLUE-RE


```
    {
        "guid": "klue-re-v1_train_00002",
        "sentence": "K리그2에서 성적 1위를 달리고 있는 광주FC는 지난 26일 한국프로축구연맹으로부터 관중 유치 성과와 마케팅 성과를 인정받아 ‘풀 스타디움상’과 ‘플러스 스타디움상’을 수상했다.",
        "subject_entity": {
            "word": "광주FC",
            "start_idx": 21,
            "end_idx": 24,
            "type": "ORG"
        },
        "object_entity": {
            "word": "한국프로축구연맹",
            "start_idx": 34,
            "end_idx": 41,
            "type": "ORG"
        },
        "label": "org:member_of",
        "source": "wikitree"
    },
```


4) 개체명 인식

- NER(Named Entity Recognition)
  - 텍스트의 개체들을 미리 정의된 카테고리(인명, 지명, 시간 등)로 분류
  - 자연어 안에서 정보를 추출해내는 task중에 하나이며, 추출된 정보들은 자연어 처리를 이용한 정보 검색과 요약, 질문 답변, 지식 베이스 구축 등 다방면에 사용됨
  - 특히 기계 번역(Machine Translation, MT)의 품질을 높이며, 사용자에게 맞춤형 번역을 제공할 수 있도록 도와주는 역할

ex) 'TWIGFARM'을 글자 그대로 해석하면 ‘트위그팜’이 아닌 ‘나뭇가지 농장’이라고 해석됨. TWIGFARM을 회사명으로 제대로 인식할 수 있다면 회사명으로써 번역될 것

- 데이터셋 예시
  - 네이버 x 창원대 NER 데이터

```
0       나는    -
1       창원대학교에서  LOC_B
2       열린    -
3       대동제를 EVT_B
4       구경하러 -
5       갔다.   -
```

5) 감성 분석

- **SA(Sentiment Analysis)**
  - 텍스트에서 정보를 추출하여 특정 주제에 대한 주관적인 인상, 감정, 태도 등을 파악하는 것
  - 고객 피드백, 콜센터 메시지 등과 같은 데이터를 분석, 기업과 관련된 뉴스나 SNS 홍보물 등에 달린 댓글의 긍/부정을 판단 등에 사용됨
- 데이터셋 예시
  - NAVER Sentiment Movie Corpus

```
id      document        label
9976970 아 더빙.. 진짜 짜증나네요 목소리       0
3819312 흠...포스터보고 초딩영화줄....오버연기조차 가볍지 않구나       1
10265843        너무재밋었다그래서보는것을추천한다      0
9045019 교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정       0
7797314 원작의 긴장감을 제대로 살려내지못했다.  0
7156791 액션이 없는데도 재미 있는 몇안되는 영화 1
```

6) 주제 분류

- **TC(Topic Classification)**
  - 주어진 텍스트에 대한 주제를 파악하여 주제별로 분류하는 것
  - 텍스트의 토픽 예측은 NLP 분야에서 핵심적인 기능으로 볼 수 있는데, 주로 뉴스처럼 이미 카테고리가 정해진 데이터 셋을 통해 구축됨
- 데이터셋 예시
  - KLUE TC


```
{
    "guid": "ynat-v1_train_00000",
    "title": "유튜브 내달 2일까지 크리에이터 지원 공간 운영",
    "predefined_news_category": "IT과학",
    "label": "생활문화",
    "annotations": {
        "annotators": [
            "08",
            "13",
            "07"
        ],
        "annotations": {
            "first-scope": [
                "생활문화",
                "생활문화",
                "IT과학"
            ],
            "second-scope": [
                "IT과학",
                "해당없음",
                "해당없음"
            ],
            "third-scope": [
                "해당없음",
                "해당없음",
                "해당없음"
            ]
        }
    },
    "url": "https://news.naver.com/main/read.nhn?mode=LS2D&mid=shm&sid1=105&sid2=227&oid=001&aid=0008508947",
    "date": "2016.06.30. 오전 10:36"
},
```