LangCon 2019

문장 속 단어

서울대학교 국어국문학과 고동현

kohdongh@gmail.com

자연어처리를 '잘' 한다는 것은?

- 인간의 자연어와 가장 가깝게 분석하기
  - 의미를 잘 살려야 한다
  - 가장 기본이 되는 것은 문장 속에서 형태소 분석하기

- "문재인 대통령"을 "문재(文才)+인 대통령"으로??
  - 이름이 '문재인'이라는 사람이 대통령이라는 의미일까
  - 글을 짓거나 글씨를 쓰는 재능이 뛰어난 대통령이라는 의미일까
  - 사람은 금방 해결할 문제이지만, 컴퓨터는 조금 고민할 문제

- "버티다 보니 살아지겠더라고요."는 어떻게?
  - 원어민의 직관만으로는 해결하기 어려운 영역도 있다

  ⇒ 언어학적 지식이 필요하다

자연어?

- 일반 사회에서 자연히 발생해서 쓰이는 언어
  예: 한국어, 영어, 일본어, 중국어 등

자연어의 특징

- 자연어의 용도: 의사소통(communication)
  - 무엇이 소통되는가?
    - 생각(thought), 가정(assumption), 정보(information) …
    - 어쨌든, 머릿속 세상을 다른 사람과 공유하려는 것
    - “오늘 아침에 어제 남은 치킨을 데워 먹었어.” / “기분이 좋아.” / “집에 가.”
    - 내 머릿속 세상을 꺼내서 보여주면 간단한데, 문제가 있다.
- 자연어의 한계 = 차원의(dimensional) 한계
  - 우리가 인지하는 세상은 3차원(공간)+시간의 흐름
  - 반면 언어는 선형적(linear)
  - 공간에 얹혀 있는 다양한 객체(objects)를 어떻게 선 위에 표현할 것인가?
    공간에서 벌어지는 복잡한 사태(situations)를 어떻게 선으로 표현할 것인가?
    = 차원의 격차를 어떻게 메꿀 것인가?

자연어의 특징

자연어의 특징

- 차원 격차를 메꾸는 전략
  - 적당히 말하고 적당히 알아듣는다
    Ex. '앨리스가 친구와 함께 길을 건넜어.'
    '앨리스가 좌우를 살피고, 길을 건넜어.' (시간순)
    '앨리스는 금발 머리이고, 검은 옷을 입었어.' (병렬)
    ⇨ 화용론(pragmatics)
  - 내 머릿속 세상의 등장인물들과 그 동작, 상태에 이름표(marker)를 달아 표현한다
    Ex. '앨리스가 횡단보도에서 파란불에 좌우를 살피며 길을 건너고 있었다.'
    ⇨ 통사론(syntax)
  - 등장인물과 동작, 상태를 개념화해서 표현한다
    Ex. 앨리스, 횡단보도, 파란불, 좌우, 살피다, 길, 건너다
    ⇨ 형태론(morphology)

문장의 이해

- “살아지겠더라고요.”

- 통사론은 문장에서 어떤 부분이 ‘이름표’인가 포착
  - 행위에 무언가 ‘아지겠더라고요’라는 이름표가 붙어 있군.
  - ‘-아지-‘라는 이름표를 보니 피동이고, ‘-겠-‘이 있으니 추측하는 것이고, …

- 형태론은 문장에서 어떤 부분이 ‘어휘’인가 포착
  - ‘살다(to live)’라는 행위를 나타내는 어휘가 들어가 있군.
  - 생명을 지닌다, 성질이나 기운이 넘친다, 물체가 제 기능을 하다, 등등

- 두 포착 작업은 거의 동시에 일어남
  - 문장에서 단어들을 분절해내는 작업이라고 할 수 있다.
  - 우리는 ‘단어‘의 정의를 조금 넓게 하는 것이 좋다

  ⇒ 통사론은 문법적 단어, 형태론은 어휘적 단어를 분절해낸다

통사론(syntax)

- 내 머릿속 세상의 등장인물들과 동작, 상태에 이름표를 달아 표현한다
  - 등장인물 역할에 따라 이름표 부여
    - 어순(word order): SVO 언어, SOV 언어 등등 (구조)
    - 조사: 이/가(주인공), 을/를(주인공에게 당하는 역할) 등등
  - 시간도 좀 표시를 해보자. 주인공이 뭘 하는데 과거, 현재, 미래 중 언제 하는지.
    - 시제(tense): V-ed, -았/었-(과거); will V, -겠/을 것-(미래) 등
  - 이야기가 진행되는 양상에도 이름표를 달아주자. 진행 중인지, 완료됐는지 …
    - 상(aspect): be -ing, -고 있-(진행상), have p.p.(현재완료), -어 있-(완료) 등
  - 내가 생각하는 이야기의 신빙성이나 의무적인 느낌 등에도 이름표를 달아주자.
    - 양태(modality): must, can, may, -더-, -구나, -네 등등
  - 나는 이 정보를 듣는 사람과의 관계도 중요해. 그 관계에도 이름표를 붙일래.
    - 높임말(honorifics): -해, -해요, -하십시오, -하오, -하게, -해라 등

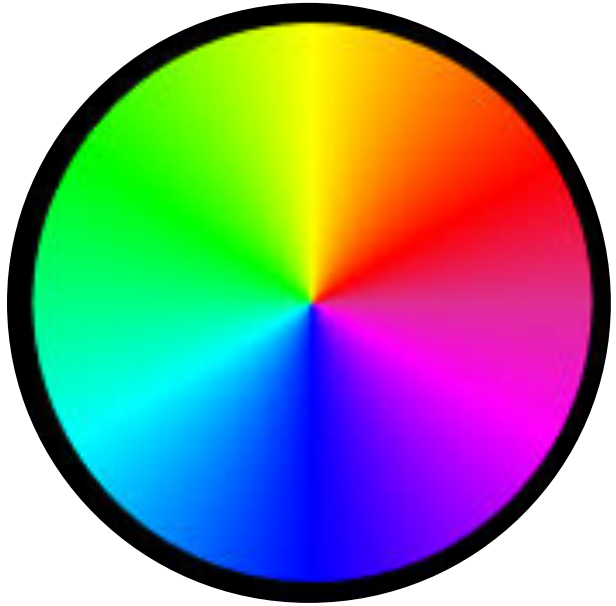
통사론(syntax)

- 이름표 달기, 즉 문법을 통해 사태를 간결하게 나타낼 수 있다
  - 머릿속 세상을 무대에 올릴 때, 등장인물들이 어떤 역할인지, 그들의 행동이나 상태는 언제 어떻게 일어났는지 등을 간결하게 나타냄.
  - 말하는 사람은 선형적인 언어를 통해, 등장인물들이 살아 움직이는 무대를 그려볼 수 있게 됨.
  - 듣는 사람은 상대방이 공유하고 싶어하는 세상을 떠올려 볼 수 있게 됨.

- 문법적인 기능만을 담당하는 요소 = 문법적 단어
  - 조사, 어미 등
  - 자기 혼자서는 쓰일 수 없음
  - 모든 자연어는 문법적 단어를 가지고 있음

형태론(morphology)

- 등장인물, 동작과 상태를 개념화한다
  - 등장인물이 너무 많은 드라마? 이해 안 된다
  - 세상에는 다양한 존재들이 있는데, 이들을 다 따로따로 나타낼 수는 없다.
  - 결국 어떤 녀석들은 같은 말로, 어떤 녀석들은 다른 말로 부를지를 결정해야 한다



- 언어마다 그 결정은 다르다
  - 한국어: '푸르다' ⊃ BLUE, GREEN
  - 영어: 'blue'; 러시아어: 'goluboy' - 'siniy'

- 세상에 존재하는 것들을 개념화해서 표현하는 요소 = 어휘적 단어
  - 우리가 흔히 생각하는 '단어(word)'
  - 문장에서는 주로 문법적 단어와 결합하여 특정한 역할을 수행함.

형태론(morphology)

- 단어끼리 붙여서 새로운 단어를 만들 수 있음 = 합성
  - 짙다+푸르다 → 짙푸르다
  - 책+가방 → 책가방

- 단어라고 하긴 조금 어려운 무언가, '접사'를 덧붙일 수도 있음 = 파생
  - 새+파랗다 → 새파랗다
  - 조심+스럽다 → 조심스럽다

- 합성인지, 파생인지, 뭣도 아닌지 헷갈리는 단어도 있음
  - 좌우, 대학교, 우물쭈물, 깝죽깝죽

통사론과 형태론, 분절의 문제

- 문법적 단어, 어휘적 단어를 어떻게 분절할 것인가?
  - ‘이효리스럽-’은 ‘이효리‘+’스럽-’로 나누고 싶은데,
    ‘상스럽-’은 나누고 싶지 않다.
  - ‘대안학교’는 ‘대안’+’학교‘로 나누고 싶은데, ‘대학교’는 나누고 싶지 않다
  - 어떻게 분절하는 것이 자연어처리에 더 유리한가?

- 기준의 설정
  - 두 요소가 독립적으로 의미가 잘 파악이 된다면 나누고,
    묶여야 정확한 의미를 파악할 수 있다면 나누지 않는다
  - 독립적인 출현 빈도가 높다면 나누고,
    같이 나오는 빈도가 높다면 나누지 않는다

통사론과 형태론, 분절의 문제

- “버티다 보니 살아지겠더라고요.” ⇒ 살+아지+겠+더라+고+요
  - 살-: 주인공이 하려는 행동을 개념화 한 것. To live.
  - -아지-: 행동을 당하는 것임을 나타내는 이름표. (피동화)
  - -겠-: 행동을 할거라고 추측된다는 이름표 (양태)
  - -더라: 일전에 경험해보니 이러이러한 행동이 있었다라는 이름표. (양태-증거성)
  - 고: 남의 말을 인용한 것처럼 이야기하고 싶다는 이름표. (인용 조사)
  - 요: 청자와 화자의 관계가 적당히 높여줘야 하는 관계라는 이름표. (높임말)

한국어 형태소 분석기의 사례

- 서울대학교 박진호 교수
- 한국어 형태소 분석은 분절(segmentation)이 필수 선결과제.
  - 영어는 단어별로 띄어쓰기가 되어 있지만, 한국어에서는 어절 단위 안에서 체언+조사, 용언+어미가 붙어 있음.
  - 이들을 분리해내는 분절 작업과 동시에 품사태깅(tagging, or labeling)이 이루어지는 것이 대부분
  - 형태소 분석을 잘해야 추후 통사 분석, 대화 분석 등의 작업에 오류가 없음

eg) '상스럽-'만을 학습해 놓고 '-스럽-'을 학습하지 않는다면,
'이효리스럽-'이 출현했을 때 제대로 대처할 수 없을 것.

  - 선형적으로 쭉 이어지는 문장 속에서 단어를 적절히 분절하는 일은 쉽지 않음
  - 대부분이 training data에서의 분절을 보고 학습이 이루어지므로, 적절한 분절을 하는 것은 연구자·개발자의 몫

한국어 형태소 분석기의 사례

- 분절 문제를 분류(classification) 문제로 전환
  - Training data에 적절한 분절이 이루어진 상태에서,
  - 어절을 형태소로 분절할 때, 어절 내 각 음절이 어떤 입출력 유형인지 분류
  - 입력 = 분절되지 않은 문장 / 출력 = 분절이 완료된 문장

- 분절의 예
  - 자주=자주: 두 음절 모두 입출력 동일 (가장 단순한 경우)
  - 먹었다=먹+었+다: 입출력은 동일하나, 형태소 경계가 삽입되어야 함
  - 갔다=가+았+다: 입력의 한 음절이 출력의 두 음절에 대응, 그 사이에 경계도.
  - 흘렀다=흐르+었+다: 입력의 두 음절이 출력의 세 음절에 대응, 그 사이 경계.
  - 더워=덥+어: 입출력 음절수는 동일, 대응하는 음절은 다름. 경계도 존재.
  - 해서=하+아서: '하다'의 불규칙 활용, 음절 개수가 줄어듦.
  - 학교랬어=학교+이+라+하+았+어: '~이라 하+었(과거)+어(반말체)'

한국어 형태소 분석기의 사례

- 입출력에서 일어나는 음절 변화를 200가지로 유형화
  - 0: 입력과 출력 음절 동일하고 경계 없음. '자0주0'='자주'
  - 1: 입출력 음절은 동일하나 뒤에 경계가 있음. '먹1어0'='먹+어'
  - 2: ㅂ 불규칙 어간, 8: ㅂ 불규칙 어미 '더2워8'='덥+어', '도2와8'='돕+아'
  - 199: 하 축약. '탄199'='하+다+ㄴ'
  - 분석기는 읽고 있는 단어가 어떤 유형에 속하는지 분류하면 됨

- LSTM 사용, 학습 데이터는 21세기 세종계획 말뭉치
- 정확도 .9653 (158,041/163,720어절)
  - 현재 가장 성능이 좋다고 알려진 U-tagger와 비슷한 수준
  - 추가적인 말뭉치를 가지고 training 과정에 있음

- 학습에 필요한 training data를 선별하는 안목
  - 좋은 데이터를 확보하는 것은 자연어처리 과정에서 필수적
  - 국립국어원 21세기 세종계획 말뭉치 역시 학습 데이터로 쓰려면 수정 필요

- 문장의 단어들을 분절하는 명확한 기준
  - 어떤 기준으로 결과를 얻어내야 할 것인가.
  - '버티다 보니 살+아지+겠+더라+고+요'
    → 어떻게 분석해야 의미를 제대로 반영할까 연구한 결과
  - 자연어처리에서, 언어학적인 지식은 이런 분절의 문제에 반드시 필요.
  - 통사 분석, 대화 분석, 자동 번역 등의 근간을 이룸.