What makes Korean NLP more difficult?

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

교착어

종류	대표적 언어	특징
교착어	한국어, 일본어, 몽골어	어간에 접사가 붙어 단어를 이루고 의미와 문법적 기능이 정해짐
굴절어	라틴어, 독일어, 러시아어	단어의 형태가 변함으로써 문법적 기능이 정해짐
고립어	영어, 중국어	어순에 따라 단어의 문법적 기능이 정해짐

교착어: 접사 추가에 따른 의미 파생

원형	피동	높임	과거	추측	전달		결과
잡						+다	잡다
잡	+히					+다	잡히다
잡	+히	+시				+다	잡히시다
잡	+히	+시	+었			+다	잡히셨다
잡			+았(었)			+다	잡았다
잡				+겠		+다	잡겠다
잡					+더라		잡더라
잡		+히	+었			+다	잡혔다
잡		+히	+었	+겠		+다	잡혔겠다
잡	+히	+었	+겠		+더라		잡혔겠더라
잡			+았(었)	+겠		+다	잡았겠다
...							...
잡	+히	+시	+았(었)	+겠	+더라		잡히시었겠더라

교착어: 유연한 단어 순서 규칙

번호	문장	정상여부
1.	나는 밥을 먹으러 간다.	O
2.	간다 나는 밥을 먹으러.	O
3.	먹으러 간다 나는 밥을.	O
4.	밥을 먹으러 간다 나는.	O
5.	나는 먹으러 간다 밥을.	O
6.	나는 간다 밥을 먹으러.	O
7.	간다 밥을 먹으러 나는.	O
8.	간다 먹으러 나는 밥을.	O
9.	먹으러 나는 밥을 간다.	X
10.	먹으러 밥을 간다 나는.	X
11.	밥을 간다 나는 먹으러.	X
12.	밥을 나는 먹으러 간다.	O
13.	나는 밥을 간다 먹으러.	X
14.	간다 나는 먹으러 밥을.	O
15.	먹으러 간다 밥을 나는.	O
16.	밥을 먹으러 나는 간다.	O

모호한 띄어쓰기

- 근대 이전까지 동양권 언어에는 띄어쓰기가 존재하지 않았음
  - 서양에서는 중세시대에 띄어쓰기가 확립됨

- 따라서, 아직 우리나라 말은 여전히 띄어쓰기와 궁합을 맞추는 중
  - 전 국립언어원장님도 어려워하시는 띄어쓰기
    참고: https://news.chosun.com/site/data/html_dir/2013/05/21/2013052103173.html

- 왜? 띄어쓰기가 어지간히 틀려도 잘 알아듣기 때문

평서문과 의문문의 차이 부재

언어	평서문	의문문
영어	I ate my lunch.	Did you have lunch?
한국어	점심 먹었어.	점심 먹었어?

주어 부재

언어	평서문	의문문
영어	I ate my lunch.	Did you have lunch?
한국어	점심 먹었어.	점심 먹었어?

한자 기반의 언어

- 표의 문자인 한자를 표음 문자인 한글로 wrapping함
  - 표의 문자: 의미 또는 사물의 형상을 글씨로 나타냄
  - 표음 문자: 사람이 말하는 소리, 음성을 글씨로 나타냄

- Wrapping 과정에서 정보의 손실 발생

茶 vs 車

단어 중의성으로 인한 문제 발생 사례

- '차'의 hidden representation

극악 난이도 한국어 NLP

- 한글은 굉장히 늦게 만들어진 문자
  - 따라서 기존 다른 문자들의 장점을 흡수
  - 굉장히 과학적으로 만들어짐

- 효율이 극대화 되었기 때문에 더욱 어려운 것

- 앞으로 우리는 자연어처리 전반 뿐만 아니라,
  한국어에 적용하였을 때의 특성과 문제도 다룰 것