



로지스틱 회귀분석과 언어모델을 이용한 내포문의 언어적 구조와 함의 관계의 연관 분석

2024 1학기 인문대학 인문아너스 심포지엄
언어학과 박유나

Q.

내포문에서 **함의 관계 판단**에 영향을 주는 언어정보를
데이터를 통해 알 수 있을까?

언어학적 개념

—

내포문 : 복문에서 하나의 절이 문장의 한 성분으로 들어간 구조

내포절 : 내포문에서 문장 성분이 되는 절

모절 : 내포절을 안은 문장

함의 : A 문장이 사실일 때 동시에 B가 사실이 되면, A가 B를 함의 한다고 함,

모순 : A 문장이 사실일 때 동시에 B가 거짓이 되면, A와 B는 모순이라고 함,

함의 관계 : 두 문장이 함의 또는 모순인지의 분류

함의 취소운용소 : 내포 명제에 대한 함의의 투사를 증명할 수 있는 언어적 기제로 부정, 의문, 조건, 양태(인식/비인식)로 유형화하였음.

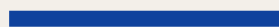
부정 : 못, 안/아니, 없, -지 말/아니하/않-, -지(는) 못하-

의문 : -까, -ㄴ가, -ㄴ데-, -나, -냐, -잖아, -지 등의 의문형 종결어미

조건 : -(으)면(야), -(ㄴ)다/라면, 어/아도

양태 : -(으)ㄴ가 싶-, -(으)ㄴ/- (으)ㄹ 것 같-, -(으)ㄹ 수 있/없-

연구 방법



Modeling

코퍼스를
이용한
회귀분석

Validation

언어모델

코퍼스명 : ‘추론_확신 분석 말뭉치 2021’ 와 ‘추론_확신 분석 말뭉치 2020’

데이터 구조와 예시

	설명	예시
idx	자료 번호	301019
source	담화 출처(파일명)	SBRW1900008528.1
genre	말뭉치 구분(문어, 신문, 구어, 대화)	구어
prev	선행 문맥	P1: 네. P5: 술을 많이 마셔도 회복이 될 겁니다. 옥희 선생님 같은 경우에도 P4: 네네 P5: 소시적에는 밤 새워서 이제 나이트에서 놀고 그러셔도 금방 회복이 되셨죠. P3: 아니 나 놀다니요. P3: 일하러 가죠. P5: 아~ 일을 하셨구나. P5: 예. P3: 예. P5: 우리가 이십대 삼십대 대체적으로 자기를 과신하고 잘못된 생활 습관 사십대 오십대 들어왔을 때 달라진 자기 몸의 환경에 맞게끔 자기 생활을 바꾸어야 되잖아요.
current	대상 문장	P5: 그렇게 바꾸어가면 만성 피로에 걸릴 일이 없거든요.
next	후행 문맥	P5: 그래서 저는 만성 피로에 걸리신 분들이 오면 반드시 이제 이거를 물어보는데요. P5: 만성 피로 뿌리 뽑는 법 땡땡땡을 높여라. P3: 면역력 P5: 면역력 P1,P2: 면역력 P5: 관계 있습니다. P1: 잠에 수면의 질을 높여라. P5: 처음부터 그렇게 정답을 맞추시면 P1: 잠깐만요. 정답이에요? all: {laughing}
pred	술어	없다
comp	보문소	르_일
ecomark	함의취소운용소	NA
eco	함의취소운용소 유형(부정, 의문, 조건, 양태, 없음)	없음
context+target	실험에서 응답자에게 제공된 담화(선행 문맥+대상 문장+후행 문맥)	우리가 이십대 삼십대 대체적으로 자기를 과신하고 잘못된 생활 습관 사십대 오십대 들어왔을 때 달라진 자기 몸의 환경에 맞게끔 자기 생활을 바꾸어야 되잖아요. 그렇게 바꾸어가면 만성 피로에 걸릴 일이 없거든요. 그래서 저는 만성 피로에 걸리신 분들이 오면 반드시 이제 이거를 물어보는데요
question	실험에서 응답자에게 제공된 질문	다음 문장에 대해 위 보기의 필자(화자)는

		어느 정도로 확신한다고 생각하십니까?
prompt	실험에서 응답자에게 제공된 내포문 기반 가설	사십대 오십대 들어왔을 때 달라진 자기 몸의 환경에 맞게끔 자기 생활을 바꾸어 가면 만성 피로에 걸린다
relation	담화 추출 작업자가 판단한 함의(확신성) 관계(함의, 중립, 모순)	모순
a1 ~ a21	응답자 응답 결과(a20, a21은 누락되었을 수 있음)	5, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 1, 2, 1, 1, 1, 1, 1, 2
95% CI lower bound	신뢰구간 하위 95% 지점	0.96404
95% CI upper bound	신뢰구간 상위 95% 지점	1.83596
mean	실험결과치 평균	1.4
sd	실험결과치 표준편차	0.994722918
entailment	실험 결과치 기반 함의(확신성) 관계 95% 신뢰 구간 기준 5~7점:함의 95% 신뢰 구간 기준 3~5점:중립 95% 신뢰 구간 기준 1~3점:모순	모순
embagr	모문-내포문의 주어 일치 여부(y:일치, n:불일치)	n
embtns	내포문의 시제(현재, 과거, 미래)	미래
embtnsm	내포문의 시제 표시	없음
general_fact	총칭문 여부	y
embper	내포문 주어의 인칭(1,2,3, 알 수 없음)	3
mtrtns	모문 시제(현재, 과거, 미래)	현재
mtrtnsm	모문 시제 표시	없음
mtrper	모문 주어의 인칭(1,2,3, 알 수 없음)	3

수집·선별 담화, 국어학 분석 정보

실험 문항, 응답 결과, 결과값 정보

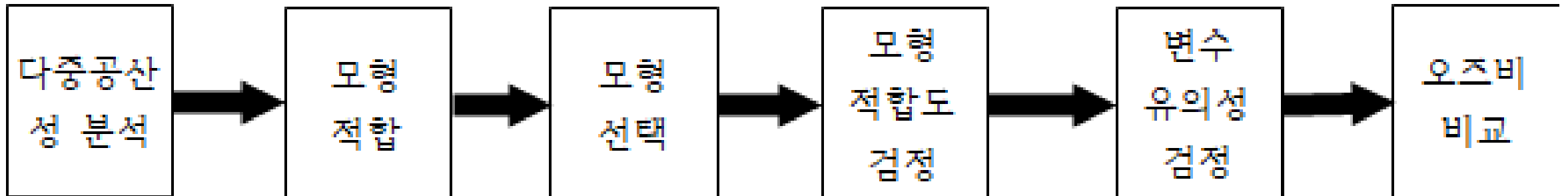
회귀 분석

회귀분석 기법 : Logistic Regression

종속 변수 : 함의 관계(함의, 모순)

독립 변수 : 함의 취소 운용소 유형, 모절과 내포절 주어의 일치 여부, 모절 시제 및 인칭,
내포절 시제 및 인칭

회귀분석 과정 도식화 :



언어모델

언어모델 : KcELECTRA

NLP TASK : Senstence Pair Classification

파인 튜닝을 위한 데이터 셋: 국립국어원 모두의 말뭉치 인공지능(AI) 말평에서 제공하는 함의 분석 평가용 말뭉치(13521)

➔ 로지스틱 회귀 분석 결과로 제시한 가설에 따라 수정한 텍스트에 대해 함의 관계가 변하는지를 확인

로지스틱 회귀분석을 통한 영향변수 도출

데이터 전처리 :

종속변수는 함의를 0, 모순을 1로 하는 이진형 변수로 재코딩

독립변수들은 독립변수들은 모두 범주형 변수이므로 더미변수로 변환

변수별 데이터 합계

함의 관계		함의 취소 운용소 유형		모절과 내포절의 주어 일치 여부			
함의	130	없음	289	일치	106		
		양태	161				
		조건	57				
모순	481	부정	52	불일치	505		
		의문	31				
		2개 이상	21				
모절 시제		모절 인칭		내포절 시제		내포절 인칭	
과거	173	1인칭	179	과거	148	1인칭	47
현재	425	2인칭	30	현재	83	2인칭	13
		3인칭	353			3인칭	548
미래	13	알 수 없음	49	미래	380	알 수 없음	3

다중 공선성 분석:

다중공산성(Multicollinearity)은 회귀분석에서 두 개 이상의 독립 변수들이 높은 상관관계를 가지는 문제
다중공산성이 존재하면 회귀 계수의 추정이 불안정해지고, 변수의 독립적인 효과를 분리하여 해석하기 어려워지
기에 모델의 설명력이 떨어진다.

모든 변수들의 값이 1과 거의 유사하므로 해당 변수들로 로지스틱 회귀분석 실시 가능

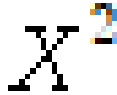
	GVIF	Df	$GVIF^{(1/(2*Df))}$
함의 취소 응용소 유형	2.24	5.00	1.08
모절과 내포절의 주어 일치 여부	1.13	1.00	1.07
내포절 시제	1.07	2.00	1.02
내포절 인칭	1.28	3.00	1.04
모절 시제	1.14	2.00	1.03
모절 인칭	2.02	3.00	1.12

모형 적합 :

AIC(Akaike Information Criterion) 비교 방식을 사용하는 Stepwise Selection을 이용한
모형 최적 모형 탐색

최종 모델 : 함의 관계 = 함의 취소 운용소 유형 + 모질의 인칭 + 모질의 시제 + 내포절과 모절
의 주어 일치 여부

모형 적합도 검정을 위한
Hosmer-Lemeshow Test 결과

	Degree of freedom	p-value
7.7927	4	0.09258

최종 모델의 로지스틱 회귀 분석 결과

	Estimate	Std. Error	z value	Pr(> z)	OR	lcl	ucl
(Intercept)	-1.141	0.197	-5.799	< 0.001	0.320	0.217	0.470
eco1	-0.244	0.323	-0.756	0.450	0.783	0.416	1.475
eco2	1.382	0.506	2.733	0.006	3.983	1.478	10.728
eco3	1.315	0.338	3.889	< 0.001	3.723	1.919	7.221
eco4	2.050	0.372	5.504	< 0.001	7.767	3.743	16.116
eco5	0.804	0.531	1.514	0.130	2.233	0.789	6.321
Subject_Equal1	-0.466	0.312	-1.492	0.136	0.627	0.340	1.157
Mat_Tense1	-1.047	0.298	-3.519	< 0.001	0.351	0.196	0.629
Mat_Tense2	0.781	0.673	1.160	0.246	2.183	0.584	8.160
Mat_Person1	-1.075	0.290	-3.709	< 0.001	0.341	0.194	0.603
Mat_Person2	-0.017	0.520	-0.032	0.975	0.984	0.355	2.727
Mat_Person3	-1.362	0.560	-2.433	0.015	0.256	0.086	0.767

Call: glm(formula = entailment ~ eco + Subject_Equal + Mat_Tense + Mat_Person, family = binomial, data = data)

각 독립변수 별 유의성 검정 :

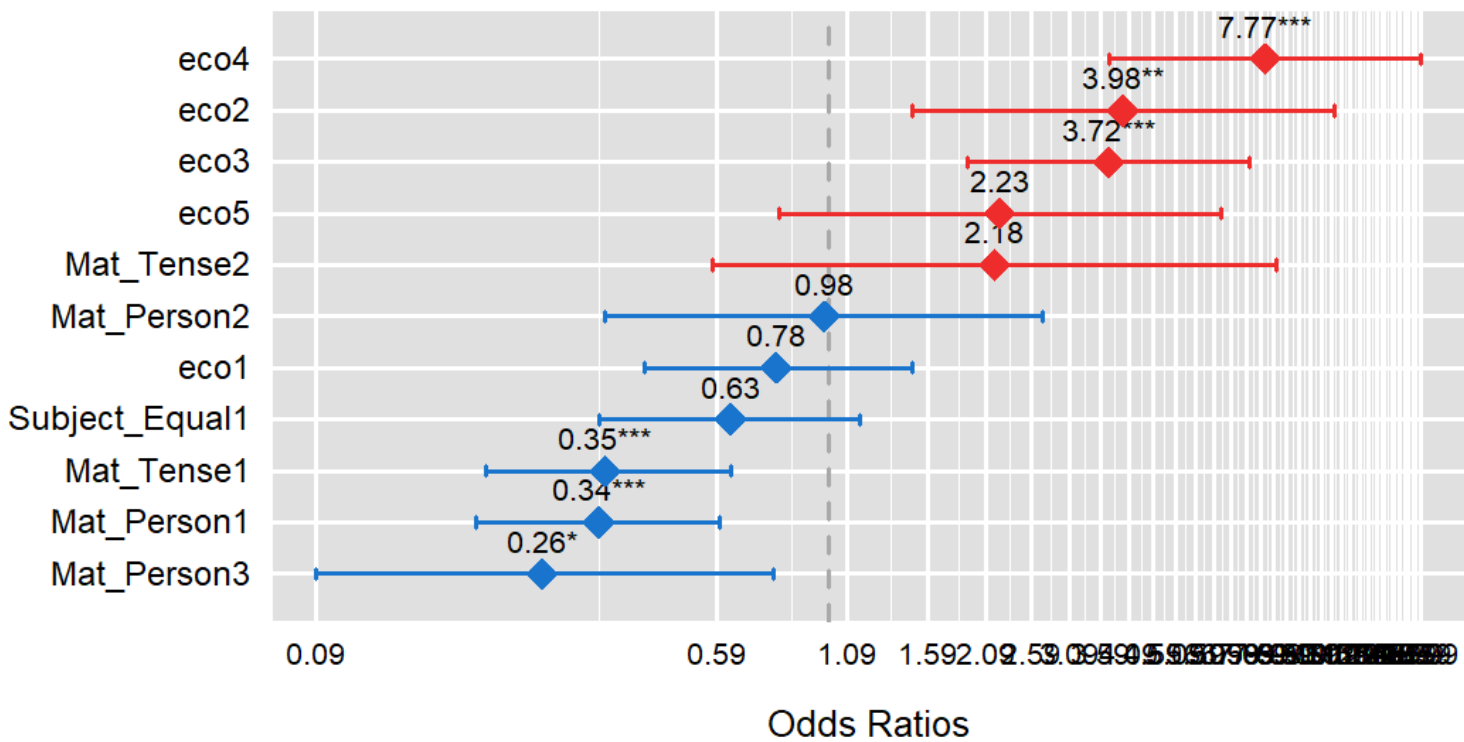
Wald 통계량을 사용 $W = \left(\frac{\beta - \beta_0}{SE(\beta)} \right)^2$

함의 최소 운용소 유형	모절과 내포절의 주어 일치 여부	모절의 시제	모절의 인칭	내포절의 시제	내포절의 인칭
3e-09	0.2022	0.0009876	0.0004265	0.115	0.6484

- ➔ 모절과 내포절의 주어 일치 여부는 최종 모델에는 포함되나 개별 변수의 유의성은 없음.
- ➔ 선행 연구 중, 영어에서 전제 함축이 존재하는 추론을 실시함에 있어 두 문장의 행위자가 일치할 경우에 추론의 정확도와 속도가 향상되었다는 선행 연구를 고려하면, 모절과 내포절의 주어 일치 여부는 함의와 모순을 분류해내는데 기여하기 보단 함의 관계 자체에 영향을 줄 수 있음. 즉 만약 내포절과 모절의 주어가 일치한다면 함의 관계가 모순인지 함의인지에 관계 없이 함의 관계를 파악하는 정확도와 속도가 향상될 수 있음.

오즈비 비교 :

오즈비는 한 변수에 대해 처치를 가했을 때의 사건의 발생 확률과 가하지 않았을 때의 사건의 발생확률에 대한 오즈의 비율. 오즈비가 높을수록 사건의 발생 확률이 높아짐.



함의 최소 운용소 유형이 조건의문부정인 경우는 함의 관계가 모순이 될 가능성을 높였다.

모절의 시제가 과거인 경우와 인칭이 1인 칭과 알 수 없는 경우는 함의 관계가 함의 가 될 가능성을 높였다.

로지스틱 회귀분석 결과에 따른 국어 내포문의 함의 관계에 대한 가설

1. 확신성 담화에 있어 함의 취소 운용소 유형 중 부정, 의문, 조건이 나타나고 모절의 시제가 현재, 인칭이 3인칭인 경우에 확신성 담화와 내포 명제가 모순 관계를 가질 가능성이 높다.
2. 확신성 담화에 있어 함의 취소 운용소 유형이 없고, 모절의 시제가 과거, 인칭이 1인칭 또는 알 수 없는 경우에 확신성 담화와 내포 명제가 함의 관계를 가질 가능성이 높다.

언어모델을 통한 가설 검증

언어모델 : KcELECTRA (구어 데이터도 학습함)

NLP TASK : Senstence Pair Classification

파인 튜닝을 위한 데이터 셋: 국립국어원 모두의 말뭉치 인공지능(AI) 말평에서 제공하는 함의 분석 평가용 말뭉치(13521)

Evaluation Metric : 정확도 0.692 / F1 0.605

1. 함의 취소 운용소가 없거나 양태인 경우 그리고 모절의 시제가 과거나 미래인 경우 그리고 인칭이 1인칭, 2인칭 그리고 알 수 없는 경우 중 함의로 분류된 데이터

➔ 함의 취소 운용소 중 부정, 의문, 조건 중 하나를 갖게 하고 모절의 시제를 현재로, 인칭을 3인칭으로 변경

결과 : 26개의 수정 데이터 중 24개를 모순으로 분류

함 의 취 소 운용소 유형	기존의 확산성 담화	수정된 확산성 담화
의문	서울에 다양한 정책 유익한 정보들 쉽게 친절하게 알려 드립니다. 친절한 과장님 순서 시작해보죠. 봄을 맞아서 서울에 유아 숲 체험원이 문을 연다는 소식 오늘 전해 드릴 수 있겠습니까. 서울시 자연 생태과에 하재호 과장과 함께 할게요.	서울에 다양한 정책 유익한 정보들 쉽게 친절하게 알려 드립니다. 친절한 과장님 순서 시작해보죠. 철수가 봄을 맞아서 서울에 유아 숲 체험원이 문을 연다는 소식 오늘 전해 드릴 수 있습니까? 서울시 자연 생태과에 하재호 과장과 함께할게요.
조건	P3: 그~ 교과 학습 세부 능력 특기사항을 통해 학생이 수업에 굉장히 열성적이면서 자기 주도적으로 참여하는 모습이 돋보였습니다. 그리고 항공기 조종사에 꿈을 이루고 싶어 하는 모습들이 자율 활동이나 진로 활동 등에 잘 나타나있었고요. 또한 학급 회장이나 학생에 임원으로서 적극적으로 학교 생활에 임하고 있음을 확인할 수 있었습니다.	P3: 그~ 교과 학습 세부 능력 특기사항을 통해 학생이 수업에 굉장히 열성적이면서 자기 주도적으로 참여하는 모습이 돋보였습니다. 그리고 항공기 조종사에 꿈을 이루고 싶어 하는 모습들이 자율 활동이나 진로 활동 등에 잘 나타나있었고요. 또한 학생이 학급 회장이나 학생에 임원으로서 적극적으로 학교 생활에 임하고 있다면 확인할 수 있습니다.
부정	촛불항쟁 때 이~ 어떤 그 뭐 이재용 구속 이런 것도 많이 외치면서 그때 저희 반올림도 방진복 입고 굉장히 열심히 이 문제가 삼성 직업병문제가 아직 해결되지 않았음을 많이 알렸었고 제가 저희가 느끼기로는 그 촛불 이후로 촛불을 거치면서 이 문제가 되게 많은 사람들에게 아~ 이게 해결되지 않았구나라는 것을 이제 인식하게 했던 상황입니다.	촛불항쟁 때 이~ 어떤 그 뭐 이재용 구속 이런 것도 많이 외치면서 그때 저희 반올림도 방진복 입고 굉장히 열심히 이 문제가 삼성 직업병문제가 아직 해결되지 않았음을 많이 알렸었고 사람들이 느끼기로는 그 촛불 이후로 촛불을 거치면서 이 문제가 되게 많은 사람들에게 아~ 이게 해결되지 않았구나라는 것을 이제 인식하지 않는 상황입니다.

2. 합의 취소 운용소가 부정, 의문, 조건, 양태인 경우 그리고 모절의 시제가 현재나 미래인 경우
그리고 인칭이 2인칭, 3인칭 그리고 알 수 없는 경우 중 모순으로 분류된 데이터

➔ 합의 취소 운용소를 없애고 모절의 시제를 과거로 인칭을 1인칭 또는 알 수 없도록 변경

모절의 인칭	기존의 확신성 담화	수정된 확신성 담화
1인칭	문제는 남북이 이 문제를 어떻게 해결하느냐이다. 정부도 북한이 천안함 사건에 대해 명시적으로 책임을 시인하거나 사과할 것으로 기대하지 않고 있다.	문제는 남북이 이 문제를 어떻게 해결하느냐이다. 나는 북한이 천안함 사건에 대해 명시적으로 책임을 시인하거나 사과할 것으로 기대했다.
알 수 없음	교육부는 각 대학·연구소의 사업단이 제출한 연구 프로젝트를 심사해 국민 세금으로 조성한 연간 1조6000억원의 막대한 연구비를 배분하고 있다. 이 때 주요 심사 기준의 하나가 교수·연구진의 논문 발표 실적이다. 교육부 장관과 교육수석부터 자기들의 논문 표절·무임승차가 문제 되지 않는다고 주장한다면 앞으로 대학교수와 연구소 박사들이 다른 사람 연구 실적을 자기 업적인 것처럼 포장해 '자격을 갖췄으니 연구비를 달라'고 할 경우 거절할 명분이 없게 된다.	교육부는 각 대학·연구소의 사업단이 제출한 연구 프로젝트를 심사해 국민 세금으로 조성한 연간 1조6000억원의 막대한 연구비를 배분하고 있다. 이 때 주요 심사 기준의 하나가 교수·연구진의 논문 발표 실적이다. 교육부 장관과 교육수석부터 자기들의 논문 표절·무임승차가 문제 되지 않는다고 주장하는 것은 앞으로 (생략) 다른 사람 연구 실적을 자기 업적인 것처럼 포장해 '자격을 갖췄으니 연구비를 달라'고 할 경우 거절할 명분이 없게 했다.

결과 : 26개의 수정 데이터 중 6개를 합의로 분류

의의 및 한계

한계 및 보완할 점:

코퍼스에 포함된 언어 정보로만 함의관계를 설명함.

➔ 술어의 사실성 여부 혹은 일반성이나 보편성을 표현하는 총칭적 표현의 포함 여부가 포함된다면 모델의 설명력이 더욱 높아질 것.

언어모델을 이용한 검증과정에서 수정 데이터의 부족. 가설검증 과정의 타당성에 대한 설명력 부족.

➔ 추가적인 확신성 판단 실험이나 수정 데이터의 타당성을 확보한 후에 가설검증을 진행하는 것이 필요할 것.

의의 :

대규모 코퍼스로 언어 현상을 모델링함

언어모델을 이용한 가설 검정이라는 새로운 방법론 이용

A modern kitchen interior featuring a light wood island with a white countertop. Two white bar stools with chrome bases are positioned in front of the island. To the left, a dark grey sofa is partially visible. The background shows a kitchen sink and a wooden cutting board on the counter. The overall lighting is soft and warm.

Q&A