로지스틱 회귀분석과 언어모델을 이용한 내포문의 언어적 구조와 함의 관계의 연관 분석

언어학과 박유나

<목차>

<초록>

본 연구의 목적은 내포문에서 함의 관계 판단에 영향을 미치는 언어 정보를 분석하는 것이다. 이를 위해 내포 명제의 진리치가 화용적인 맥락과 국어학적 요인들과 관련될 수 있음을 가정하고, 코퍼스를 기반으로 통계적 분석을 수행한다. 연구에 사용된 코퍼스는 국립국어원의 '추론_확신 분석 말뭉치 2020'과 '추론_확신 분석 말뭉치 2021'이다. 로지스틱 회귀분석과 다양한 통계적 가설 검정법을 통해 통해 내포문의 구조적 요소들이 함의 관계에 미치는 영향을 분석하였다. 분석 결과 함의 취소 운용소 유형, 모절의 시제 및 인칭, 모절과 내포절의 주어 일치 여부가 함의 관계에 영향을 주는 요인임을 확인했다. 회귀모델에 입각한 가설을 수립 후 언어모델을 활용하여 검정한 결과 함의 취소 운용소 유형 중 부정, 의문, 조건이 나타나고 모절의 시제가 현재, 인칭이 3인칭인 경우에 확신성 담화와 내포 명제가 모순 관계를 가질 가능성이 높음을 발견했다. 이를 통해 내포문의 구조적 특성과 함의 관계 판단의 연관성을 확인할 수 있었으며, 본 연구는 코퍼스 기반의 언어 분석과 통계적 방법론의 결합을 통해 언어 현상을 객관적으로 설명하고, 이후 이론적 논의로 확장할 수 있는 기반을 마련하였다.

1. 서론

1.1 연구 목적

해당 연구는 코퍼스를 기반으로 통계적 분석을 통해 내포문에서 함의 관계 판단에 영향을 주는 언어정보를 밝히는데 목적이 있다. 이를 통해 국어 내포문의 복잡한 의미론적, 문법적, 화용론적 특성을 분석하고, 내포문 구조에서의 다양한 함의 관계를 체계적으로 이해하는 데 기여하고자 한다. 통계적 분석 모델은 로지스틱 회귀분석을 사용하고 회귀분석의 타당성을 입증하기 위해 여러 통계적 가설 검정법을 사용하여 연구의 신뢰성을 확보하고자 한다.

1.2 연구 방법

코퍼스를 이용한 통계분석과 언어모델을 이용할 것이다. 연구의 단계는 요약하자면 로지스틱 회귀분석을 이용한 확신성 판단에 개입하는 언어정보 탐색, 언어모델의 문장 관계 분류 방법을 이용한 모형의 타당성 검증이다.

연구를 위해 사용하는 코퍼스는 국립국어원에서 제공하는 '추론_확신 분석 말뭉치 2020'과 '추론_확신 분석 말뭉치 2021'이다. 해당 코퍼스는 담화에 포함된 내포 명제의 사실성에 대한 화자의 직관을 언어 실험을 통해 수집한 정보와 실험자들에게 제공된 담화에 대한 언어정보로 이루어져 있다. 예를 들어, 담화 '철수는 바나나를 구황작물이라고 생각해. 바나나가 땅에서 자라는 줄 알지?'가 주어졌을 때, 내포문 '바나나가 땅에서 자란다'을 어느 정도로 확신하는지를 7점 척도를 사용하여 일반 언어 사용자에게 실험을 진행한다. 피실험자들의 점수의 평균치를 바탕으로 도출한 함의 관계가 태깅되어있다. 더불어 코퍼스에는 함의 관계에 영향을 줄 수 있으리라 생각하는 통사/의미적 언어정보(함의취소운용소, 모문-내포문의 주어일치여부, 내포문의 시제, 내포문의 주어의 인칭, 모문시제, 모문의 주어 인칭)들을 교육을 받은 작업자들에 의해 분류하여 태깅해놓은 정보도 포함되어 있다.

이후 해당 언어정보들이 화자의 함의 관계 판단에 기여하는 바를 파악하기 위해 로지스틱 회귀분석을 실시한다. 독립변수는 코퍼스에 부착된 언어정보들이며, 종속변수는 함의관계(함의, 모순)에 해당한다. 변수들의 VIF 비교, 모델의 AIC 비교, Hosmer-Lemeshow 검정, Wald 검정의 통계적 지표와 방법론을 이용하여 분석의 설득력을 높이고자 한다. 더불어 각 설명변수들의 오즈비(독립변수의 변화에 따른 종속변수의 변화 비율) 비교를 통해 종속변수에 가장 높은 상관성을 지닌 독립변수를 찾아내는 것을 목표로한다. 이를 통해 함의 관계 판단에 대한 화자의 직관에 영향을 주리라 생각되는 언어정보에 대한 가설을 세운다.

마지막은 언어모델을 이용하여 문장 분류의 정확도를 계산하여 로지스틱 회귀분석 결과로 산출한 종속변수와 독립변수간의 관계에 대한 가설을 검정하고자 한다. 로지스틱 회귀분석 결과 함의 관계에 대한 공헌도가 높은 변수에 대해 가설을 수립하고 코퍼스에서 분리한 테스트 데이터에 인위적인 조작을 가해 예측 결과를 비교한다. 예를 들어, 내포절과 모절의 주어 일치 여부가 함의 관계에 영향을 미치는 변수라면 테스트 데이터의 담화를 내포절과 모절의 주어가 일치할 수 있도록 수정한다. 그리고 언어모델에게 코퍼스를 잘 반영할 수 있도록 파인-튜닝을 진행하고 수정한 테스트 데이터에 대하여 Sentence Pair Classification를 실시하여 예측 결

과를 통해 로지스틱 회귀 분석 결과를 검증한다. Sentence Pair Classification은 대표적인 NLP 태스크 중 하나로 두 분장의 관계를 함의 또는 모순으로 분류하는 태스크를 말한다.

1.3 연구 의의

해당 연구는 개별 연구자의 언어 직관에 호소하는 것이 아닌 코퍼스를 통해 언어현상을 설명하려 시도한다는 점에서 의의가 있다. 코퍼스 언어학(Corpus Linguistics)은 자연 언어 텍스트의 집합인 코퍼스를 사용하여 언어를 연구하는 분야이다. 이 접근법은 언어 데이터베이스인 코퍼스를 분석하여 언어의 구조, 사용, 그리고 의미를 연구한다. 코퍼스는 다양한 장르와 분야에서 수집된 텍스트의 집합으로, 소설, 신문 기사, 대화 기록, 학술 논문 등 여러 유형의 텍스트가 포함될 수 있다.

코퍼스는 실제로 사용된 언어 데이터를 포함하므로, 언어 사용의 현실적이고 자연스러운 예를 제공한다. 이는 이론적 모델이나 직관에 의존하는 것보다 더 실제 언어 사용을 잘 포착하는 결과를 도출할 수 있게 된다. 코퍼스 데이터는 언어학자의 직관이나 편견에 영향을 받지 않는 객관적인 자료이다. 이는 연구 결과의 신뢰성을 높이며, 동일한 코퍼스를 사용하면 다른 연구자들도 동일한 결과를 얻을 수 재현가능성의 측면에서 장점을 갖는다. 통계적 분석을 도입함에 있어 많은 데이터를 확보하는 것은 중요하다. 코퍼스를 이용한 분석은 대규모 데이터를 효율적으로 분석할 수 있으므로 통계적 기법을 사용하기에 적절하다. 또한 소규모 데이터나 직관에 의존해서는 발견하기 어려운 언어적 특성 대규모 데이터를 통해서는 언어 사용의 미묘한 패턴이나 드문 현상을 발견할 수 있다는 장점이 있다.

또한 해당 연구는 데이터를 통한 가설 설정 후 언어모델을 이용한 가설 검정이라는 새로운 방법론을 도입하였으며, 한국어의 내포문의 함의 관계와 언어적 특성의 연관성을 통계적 방법을 통해 파악하는 것에 있으므로 이후 원인에 대한 이론적 논의로 확장할 수 있다는 점에서 의의가 있다.

2. 선행 연구

본 연구의 목적이 코퍼스를 통해 함의 관계에 영향을 미치는 언어학적 요인들을 파악하고 경향성을 발견하는데에 있다. 따라서 언어학적 이론에 입각한 원인 분석은 실시하지 않는다. 그렇지만 관련 선행 연구를 검토하는 것은 해당 연구에서 분석 대상으로 삼는 언어학적 요인들이 어째서 함의 관계에 영향을 줄 가능성 있는 변수들로 선정되었는지에 대한 이해를 증진할 수 있을 것이다.

(Waslh and Jonson 2004) 에 따르면 영어에서 전제 함축이 존재하는 추론을 실시함에 대한 두 문장의 행위자가 일치할 일치할 경우에 추론의 정확도와 속도가 향상되었다. 예를 들어, 아래와 같은 추론은 전제들로부터 결론의 함축이 일어난다. 만약 A와 B의 행위자가 일치한다면, 추론이 용이해짐을 실험적 연구를 통해 밝혔다. 비록 (Waslh and Jonson 2004)는 전제와 결론을 갖는 추론에 있어서의 함의 관계에 대한 논의였지만, 이를 내포문 구제로도 확

장하여 이해할 수 있는 여지가 존재한다. 따라서 내포문 구조에서도 모절과 내포절의 주어가 일치한다면 함의 관계를 파악하는데 도움이 될 수 있을 것이란 추측이 가능하다.

1) 전제 1 : A or B but not Both.
2) 전제 2 : A
3) 결론 : not B

(Kiparsky 1971)에 따르면 확신성과 관련된 화자의 태도는 모절의 술어가 지닌 사실성에 의해 파악된다. 사실성 술어는 내포절의 진리값이 참임을 전제하고 비사실성 술어는 내포절의 진리값이 참임을 전제하지 않는다.[1] 모문에 술어가 사실성 술어인 경우에 함의 취소 운용소인 양태, 조건, 의문 부정 등이 존재하더라도 내포절은 진리치는 보존될 수 있다. 예를 들어, 'John dosen't know that it is snowing.'이라는 문장에 대해 내포절인 'it is snowing'은 모절에 부정이 존재할지라도 바뀌지 않는다. 이를 전제 투사 원리라 한다. 그러나 (de Marneffe et al.(2019)에 따르면, 내포 명제와의 함의 관계가 단순히 모절 술어의 사실성에 의존하는 것이 아니라 화용론적 상호작용도 영향을 미치기에 다양한 양상이 나타난다. 확신성 실험을 통해 여러 문장에 대해서 피실험자의 함의 관계 판단을 종합하였을 때, 비사실성 술어에 대해서도 화자가 강한 확신성을 갖는 사례들이 많았다. 이러한 판단에는 사실성 술어와 맥락에 영향을 주는 통사적 요소들, 인칭, 시제 등이 영향을 주었다.

한국어의 경우에도 마찬가지로 내포 명제의 진리치를 단순히 사실성 술어에 입각해서만이 아니라 화용적인 맥락과 국어학적 요인들과의 연관 속에서 파악할 수 있다. 예를 들어, '철수는 바나나를 구황작물이라고 생각해. 바나나가 땅에서 자라는 줄 알지?'는 내포 명제 '바나나가 땅에서 자란다'를 함의하는지 또는 내포 명제의 사실성을 취소(cancellation)하는지가 화자(청자)의 직관에 따라 매우 다양한 양상으로 나타날 수 있을 것이다.

3. 코퍼스 설명

본 연구에 사용되는 코퍼스는 국립국어원에서 제공하는 '추론_확신 분석 말뭉치 2021'과 '추론_확신 분석 말뭉치 2020'이다. 해당 코퍼스는 확신성 담화에 포함된 내포 명제의 사실성에 대한 화자의 직관을 언어 실험을 통해 수집한 정보와 실험자들에게 제공된 담화에 대한 언어 정보로 이루어져 있다. 분석 결과를 제시함에 앞서 코퍼스의 구성 과정과 구성 요소에 대해 소개함으로써 분석 대상에 대한 이해를 증진하고 본 연구의 문제의식을 분명히 하고자 한다.

3.1 담화 구성

확신성 담화는 어떤 대화 또는 발화가 함의하고 있는 내용의 사실성이 국어의 문장구조와 맥락에 의하여 결정되는 담화를 말하며, 국립국어원 신문기사, 문어, (준)구어 말뭉치를 대상으

1) 사실성 술어는 see, find, know, realize, notice 등이 해당하고 비사실성 술어는 occur, suppose, seem, assume, think, believe 등이 해당한다.

로 내포 명제를 포함하고 있는 문장을 앞뒤의 문맥과 함께 추출되었다. 내포문 복문에서 하나의 절이 문장의 한 성분으로 들어간 구조로 이때의 문장 성분이 되는 절을 내포 명제라 한다. 예를 들어, '철수가 영희를 역 앞에서 보았다고 말했다'는 문장에 대해 '철수가 영희를 역 앞에서 보았다'가 내포 명제가 되며, 해당 문장을 내포 명제를 함의(entailment)하고 있다. 내포 명제는 화자의 확신성에 대한 판단의 대상이 되는데 그 이유는 내포 명제와 전체 문장 간에 함의 관계가 존재하기 때문이다. A가 B를 함의한다는 것은 A 문장이 사실이면 동시에 B가 사실이 됨을 말한다. 즉 종전의 문장들이 함의 관계에 있으므로 화자가 내포 명제를 사실로 생각한다는 직관을 담고 있다고 볼 수 있다. 확실성 담화의 경우 구어, 문어 등 다양한 코퍼스에서 추출된 것이므로 맞춤법과 띄어쓰기, 기호사용 오류, 주어와 술어 및 다른 문장 성분의 생략이 나타나기도 한다. 따라서 확신성 담화에서 내포 명제를 추출하는 과정에서 수정과 복원이 발생하였다. 내포명제에 맞춤법과 띄어쓰기 그리고 기호 사용 등의 오류는 수정하였으며, 내포절에 주어가 없어 명제의 내용을 파악하기 어려운 경우엔 선행 주어를 복원하고 인칭대명사의 선행어를 담화에서 찾을 수 있는 경우엔 선행어를 복원하는 과정을 거쳤다. 이러한 과정을 통해 구성된 확실성 담화와 내포 명제 그리고 둘의 관계에 대한 예시는 <표 1>과 같다.

장르	확실성 담화	내포 명제	추론 관계
구어	짠맛에 대명사라 할 정도로 짜디짠 된장이지만 오백 년 전에 만들었던 장은 염도가 어떻게 될까? 일정한 양을 물에 희석해서 한번 염도를 측정해 봤더니요. 평균 무려 십 점 이오 퍼센트로 일반 된장보다 염도가 낮은 것으로 확인이 됐습니다. 이렇게 오백 년 전통장 맛은 어떨까요?	오백 년 전에 만들었던 장은 일반 된장보다 염도가 낮다	함의
신문	행정기관 전산망에는 쌍둥이 모자가 영암 모 아파트에 거주한다고 나왔으나, 현장 조사를 해보니 애초부터 이 곳에 살지 않았다는 주변인들의 진술을 얻는 데 그친 것이다. 전남도교육청은 쌍둥이 친모가 아들들을 데리고 출국했을 가능성을 고려해 출입국사무소에 문의했으나 출국기록은 없었다. 지난 17일 교육당국 의뢰로 경찰 수사가 개시되면서 상황은 예사롭지 않게 흘러갔다.	쌍둥이 친모가 아들들을 데리고 출국했다	모순

표 1. 담화 사례

본 연구에서는 더 많은 데이터 확보를 위해 '추론_확신 분석 말뭉치 2021'과 '추론_확신 분석 말뭉치 2020'을 모두 사용하였지만 20년도와 21년도의 담화 구축 방식에 있어서 차이가 존재한다. 20년도의 경우에 모든 담화는 함의 취소 운용소를 포함해야 했던 반면, 21년도에는 함의 취소 운용소가 존재하지 않은 담화도 수집 대상에 포함되었다. 또 확신성 담화의 술어구조에 있어 20년도에 술어 구조의 보문소는 '-음/ㅁ, -기, 것을, 것으로, -다고/라고, 줄'의 6가지로 제한되었고 모문의 술어는 사실성을 기준으로 사실성, 비사실성, 반사실성으로 구분가능한 단어들만으로 제한되었다. 반면 21년도에는 술어의 유형을 제한하지 않고 넓은 맥락을 고려하여 맥락이 내포 명제에 대한 확신성에 영향을 미치는 경우까지 반영하였기에 형식이 다양한 담화를 포함하고 있다. 데이터의 무결성에 있어 데이터간의 관계에 대한 적절성을 유지하는 것은 중요하지만, 주요한 분석 대상인 확신성에 대한 담화임에는 틀림없다는 점과 데이터

를 결합할 경우 규격화된 정보만을 포함하지 않고 다양한 구성을 가지기에 실제 언어 사용을 더 잘 반영할 수 있다는 점에서 연구의 주요한 주제의식을 벗어나지 않을 것이란 판단하에 '추론_확신 분석 말뭉치 2021'과 '추론_확신 분석 말뭉치 2020'을 모두 사용하였다.

3.2 확신성 언어 실험

'추론_확신성 분석 말뭉치'에는 일반 언어 사용자 대상 실험을 통해 내포문을 기반으로 생성한 가설에 대하여 화자가 확신하는 정도를 점수화한 실험에 대한 정보가 포함되어 있다.

실험은 21년도에는 20명, 20년도에는 평균 8명의 한국어가 모어인 만 19세 이상 남녀를 대상으로 온라인 설문으로 진행되었다. 실험 참여자는 확신성 담화를 읽고 필자(화자)가 내포 명제를 어느정도로 확신하는지에 대한 질문에 7점 척도로 답한다. 해당 점수치를 품질 일관성 확보를 위한 통계적 검정을 거친 후에 평균치를 기준으로 함의 관계 정보를 부착한다. 이때 21년도에는 함의, 중립, 모순으로 구분하였지만 20년도에는 함의와 모순으로만 구분했다는 차이점이 있다. 해당 연구에서는 21년도의 말뭉치에 중립으로 분류된 데이터가 365개중 39개뿐이기에 중립으로 구분된 데이터를 삭제하고 함의와 모순의 관계만을 갖는 데이터를 사용했음을 밝힌다.

확신성 언어 실험은 문장의 함의 관계를 담화수집자의 직관적 판단에 호소하지 않고 일반 언어 사용자의 통계치를 사용함으로써 한국어 화자의 일반적인 직관을 계량화했다는 점에 의의가 있다.

3.3 언어 주석 정보

'추론_확신 분석 말뭉치'에는 함의 관계에 영향을 줄 수 있는 통사/의미 이론적 언어 정보가 부착되어 있다. 코퍼스에 포함된 언어 정보는 보문소, 모절 술어, 함의 취소 운용소, 내포절과 모절의 시제정보와 주어의 인칭, 내포절과 모절의 주어 일치 여부이다. 21년도 코퍼스에는 총칭문 여부가 20년도 자료에는 사실성에 대한 정보가 포함되어 있다.

	내포절				모절			
시제	과거		현재	미래	과거		현재	미래
시제소	은/는	었/았	(으)ㄹ-것/겠	없음	은/는	었/았	(으)ㄹ-것/겠	없음
인칭	1인칭	2인칭	3인칭	알 수 없음	1인칭	2인칭	3인칭	알 수 없음

표 2 내포절과 모절의 언어 정보

보문소는 내포절이 전체 문장의 보문이 되도록 만드는 요소로 'ㅁ을', '것으로' 등이 해당한다. 함의 취소 운용소는 내포 명제에 대한 함의의 투사를 증명할 수 있는 언어적 기제로 부정, 의문, 조건, 양태(인식/비인식)[2]로 유형화되어 있다. 내포절과 모절에 대한 언어정보의 세부적인

2) 각 함의 취소 운용소에 대한 예시는 다음과 같다. 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-,

사항은 <표 2>와 같다.

4. 로지스틱 회귀분석을 통한 영향변수 도출

코퍼스에 포함된 언어정보들이 함의 관계 판단에 기여하는 바를 파악하기 위해 로지스틱 회귀분석을 실시한다. 회귀분석은 여러 변수간의 관계를 분석하여 종속 변수를 독립변수들로부터 설명하고 예측하는 통계적 기법이고 로지스틱 회귀분석은 종속 변수가 이진형인 경우에 사용된다. 즉, 종속 변수가 두 가지 범주(예: 성공/실패, 합격/불합격, 참/거짓) 중 하나를 가질 때 사용된다. 로지스틱 회귀분석은 로지스틱 함수를 통하여 종속 변수에 대한 예측확률을 구할 수 있다.

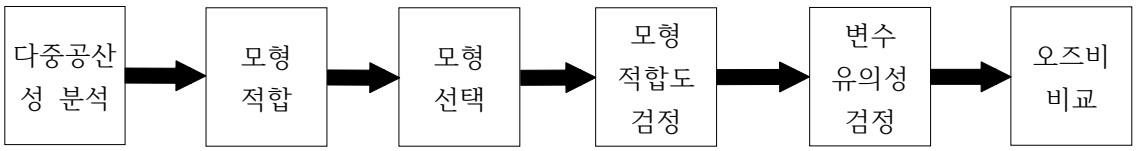

그림 2 로지스틱 회귀분석 과정

독립변수는 코퍼스에 부착된 언어정보들이며, 종속변수는 함의 관계(함의, 중립)에 해당한다. 종속변수와 독립변수가 모두 범주형 변수이기에 로지스틱 회귀분석을 이용한다. 분석 과정은 다음과 같다. 먼저, 다중공산성 분석 후 로지스틱 회귀 모형을 적합한다. 그 후 Stepwise selection 방식으로 AIC(Akaike information criterion) 비교를 통해 함의 관계에 대한 최적 모형을 선정하고 모형의 적합도(good of fit)을 검정한다. 마지막으로 각 독립변수들의 설명력을 확인하기 위해 Wald 검정과 오즈비(독립변수의 변화에 따른 종속변수의 변화 비율) 비교를 통해 종속변수에 가장 높은 상관성을 지닌 독립변수를 찾아내는 것을 목표로 한다.

4.1 데이터 전처리

함의 관계에 영향을 미치는 언어학적 요인 분석을 진행하기에 앞서 전처리 과정을 다음과 같이 실시하였다. '추론_확신 분석 말뭉치 2021'과 '추론_확신 분석 말뭉치 2020'을 합하여 640개의 데이터를 확보하였으며, 이 중 함의 관계가 중립으로 분류된 39개를 삭제하고 총 611개의 데이터를 이용하여 분석을 실시하였다. 코퍼스에 부착된 언어 정보 주석 중 정량화가 어려운 술어, 보문소, 함의취소운용소, 모절과 내포절 시제 표지 등을 제외하고 함의 취소 운용소 유형, 모절과 내포절 주어의 일치 여부, 모절 시제 및 인칭, 내포절 시제 및 인칭을 독립변수로 설정하였다. 함의 취소 운용소 유형의 경우 분류를 위해 취소 운용소가 2개 이상이 있는 경우를 '2개 이상'이라는 카테고리를 따로 만들어 그룹화하였다. 종속변수로는 함의 관계를 사용하였는데, 이는 2.2절에서 서술한 언어 실험에 따른 결과로 일반 사용자의 언어 직관을 반영한다. 각 변수별 범주의 합계는 <표 3>과 같다.

-(으)ㄹ 텐데, -(으)ㄹ지 모르-, -겠(추측)-, -(으)ㄹ 필요가 있-, -(으)면 되-, -(으)ㄹ 수 있/없- ,겠-, -고 싶-, -고자 하-

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

표 3 변수별 합계

종속변수는 함의를 0, 모순을 1로 하는 이진형 변수로 재코딩해 주었다. 독립변수들은 독립변수들은 모두 범주형 변수이므로 회귀분석을 진행하기 위해서는 연속형 변수처럼 바꿔어주어야하고 이를 더미변수로의 변환이라 한다. 각 변수마다 범주를 구분하기 위해 더미변수를 생성하고 자신이 속한 집단에 1, 이외의 집단은 0으로 변수를 정한다. N개의 집단을 분석하는 경우 선형독립성 보장을 위해 N-1개의 더미변수가 분석에 요구되며, 한 범주는 기준집단(reference group)으로 지정해야한다. 기준집단은 다른 더미변수들과의 관계를 설명하며, 각 범주가 독립변수과 종속 변수 간에 어떠한 영향을 미치는지에 대한 정량적인 분석이 가능하게 한다.

기준집단의 경우에는 데이터가 많은 범주를 선정하는 것이 일반적이기에 함의 취소 운용소 유형은 없음을, 모절과 내포절의 주어 일치 여부는 불일치를, 모절과 내포절의 시제의 경우에는 현재를, 모절과 내포절의 인칭의 경우에는 3인칭을 기준집단으로 선정하였다.

4.2 모형 구축 및 결과 분석

4.2.1 다중공산성 분석

다중공선성(Multicollinearity)은 회귀분석에서 두 개 이상의 독립 변수들이 높은 상관관계를 가지는 문제를 말한다. 즉, 독립변수들간에 연관성이 있는 경우를 말한다. 다중공선성이 존재하면 회귀 계수의 추정이 불안정해지고, 변수의 독립적인 효과를 분리하여 해석하기 어려워지기에 모델의 설명력이 떨어진다. 따라서 모형을 수립하기 전에 변수들간의 다중공산성이 있는지를 확인해야 한다.

다중공산성을 파악하기 위해서는 VIF(Variance Inflation Factor)를 활용할 수 있다. VIF는 특정 독립 변수가 다른 독립 변수들로 인해 얼마나 분산이 증가하는지를 나타낸다. VIF 값이 높을수록 해당 독립 변수는 다중공선성의 문제가 심각하다는 의미이다.

	GVIF	Df	GVIF^(1/(2*Df))
함의 취소 웅용소 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

표 4 독립변수의 VIF

주목해야 할 값은 GVIF^(1/2*DF)로 다중 범주형 변수를 갖는 데이터에 대해, VIF값을 독립변수의 수로 나누어 변수별 스케일을 맞춘 값으로 1과 비슷하면 다중공산성이 거의 없다고 볼 수 있다. 모든 변수들의 값이 1과 거의 유사하므로 해당 변수들로 로지스틱 회귀분석을 실시해도 무방하다.

4.2.2 모형 적합

<표 5>는 모든 설명변수에 대한 로지스틱 회귀분석 결과이다. 각 항목들이 의미하는 바를 기술하기에 앞서 모델 선택 과정이 필요하다. 모델 선택은 여러 후보 모델 중에서 최적의 모델을 선택하는 과정을 말한다. 해당 과정이 필요한 이유는 모든 변수를 포함한 모델이 항상 최적의 예측 성능을 가지는 것은 아니기 때문이다. 모델의 예측 성능을 저하시키는 불필요한 변수를 제거하는 모델 선택을 통해 최적의 변수 조합을 찾으면, 훈련데이터에 지나치게 맞춰져 새로운 데이터에 대해 일반화 능력이 떨어지는 오버피팅을 방지하고 실제 데이터에 대한 예측력을 확보하여 모델의 설명력을 높일 수 있다. 또 모델에 포함된 변수가 많아질수록 모델의 해석이 어려워지기 때문에 모델 선택을 실시한다. 변수 선택을 통해 중요한 변수만 포함된 간결한 모델을 만들면, 변수 간의 관계를 더 명확하게 이해할 수 있다.

모델 선택을 위한 여러 방법중 Stepwise Selection을 사용한다. 이는 Forward Selection과 Backward Elimination을 결합한 방법으로 변수들을 추가하거나 제거하며 AIC(Akaike Information Criterion) 비교를 통해 최적의 모델을 찾는 방법이다, AIC는 모델의 적합도와 복잡도를 동시에 고려하여 최적의 모델을 선택하는 지표이다.

$$AIC = 2k - 2\log(L)$$

AIC의 정의는 위와 같으며 k는 모델에 포함된 변수의 수로 모델의 복잡도를 반영하는 패널티텀이다. log(L)은 모델의 최대 우도(log-likelihood)이다. 우도는 모델이 관측된 데이터를 얼마나 잘 설명하는지를 나타내는 척도이다. 우도를 이해하기 위해서는 회귀모델의 수식에 대해 살펴볼 필요가 있다. 독립변수들로 종속변수를 설명하는 확률값을 산출해내는 수식은 아래와 같다. 독립변수(x)와 회귀계수(β)의 선형결합(1)은 사건의 발생확률의 로짓값(2)이다. 로짓함수는 사건이 발생하지 않을 확률과 사건이 발생할 확률의 비율에 로그를 취하는 변환 함수(2)이고 이를 역함수를 취해주면 확률값(3)을 구할 수 있다. n개의 관측치가 있다고 가정하면, 사건

	Estimate	Std. Error	z value	Pr(>|z|)	OR	lcl	ucl
(Intercept)	-1.180	0.221	-5.331	< 0.001	0.307	0.199	0.474
eco1	-0.198	0.328	-0.605	0.545	0.820	0.431	1.559
eco2	1.403	0.508	2.765	0.006	4.069	1.505	11.002
eco3	1.283	0.347	3.693	< 0.001	3.608	1.826	7.130
eco4	2.062	0.380	5.432	< 0.001	7.861	3.736	16.543
eco5	0.773	0.553	1.399	0.162	2.166	0.734	6.398
Subject_Equal1	-0.408	0.320	-1.277	0.202	0.665	0.355	1.244
Inner_Tense1	-0.177	0.286	-0.620	0.536	0.838	0.478	1.468
Inner_Tense2	0.478	0.303	1.579	0.114	1.613	0.891	2.921
Inner_Person1	-0.327	0.472	-0.693	0.489	0.721	0.286	1.818
Inner_Person2	0.670	0.668	1.003	0.316	1.954	0.528	7.231
Inner_Person3	-13.800	783.872	-0.018	0.986	0.000	0.000	Inf
Mat_Tense1	-1.038	0.301	-3.453	0.001	0.354	0.197	0.638
Mat_Tense2	0.741	0.657	1.127	0.260	2.098	0.578	7.607
Mat_Person1	-1.112	0.297	-3.750	< 0.001	0.329	0.184	0.588
Mat_Person2	-0.083	0.537	-0.155	0.877	0.920	0.321	2.635
Mat_Person3	-1.277	0.563	-2.267	0.023	0.279	0.092	0.841

Call: glm(formula = entailment ~ ., family = binomial, data = data)

표 5 모든 독립 변수에 대한 로지스틱 회귀분석 결과

이 일어날 확률은 (4)와 같고 관측된 데이터에 대해 모델이 예측한 확률을 곱하여 전체 데이터의 우도를 계산한다. 우도를 계산할 때 곱셈이 포함되어 있기에 로그를 취하여 계산을 단순화한다.(6) 식을 통해서 알 수 있는 것은 관측된 결과와 모델이 예측한 값이 일치할수록 우도가 높게 난다는 것이다. 즉 우도는 모델이 예측한 확률이 실제 관측된 결과와 얼마나 일치하는지를 나태내는 지표고 높을수록 모델이 데이터를 더 잘 설명하며 관측된 데이터와 모델이 예측한 결과 간의 차이가 적다는 것을 의미한다. AIC의 정의에 따라 우도가 높을수록 AIC가 낮아지므로 모델 선택과정에서 변수를 추가, 제거하며 AIC가 낮은 모델을 선택한다.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad \cdots (1)$$

$$z = logit(p) = \ln\left(\frac{p}{1-p}\right) \quad \cdots (2)$$

$$p = \frac{1}{1+e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k}} \quad \cdots (3)$$

$$p_i = \frac{1}{1+e^{\beta_{i0} + \beta_{i1} x_{i1} + \beta_{i2} x_{i2} + \cdots + \beta_{ik} x_{ik}}} \quad \cdots (4)$$

$$L(\beta) = \prod_{i=1}^{n} p_i^{Yi}(1-p_i)^{1-Yi} \quad \cdots (5)$$

$$\log L(\beta) = \sum_{i=1}^{n} [Y_i \log(p_i) + (1-Y_i)\log(1-p_i)] \cdots (6)$$

Stpewise Selection을 통해 최종 선택된 모델을 축소모델이라하며 축소 모델에 대한 회귀 분석결과는 <표 6>에서 확인할 수 있다. 축소모델에서는 내포절의 시제와 인칭이 제거되었다. eco는 함의 취소 운용소(Entailment Cancelation Operator)의 약자로 함의 취소 운용소가 없을 때를 기준그룹으로 1~5는 각 양태, 조건, 부정, 의문, 2개 이상에 해당한다. Subject_Equal1은 내포절과 모절의 주어 일치 여부로 불일치를 기준 그룹으로 하므로 주어가 일치할 때를 말한다. Mat_Tense와 Mat_Person은 각각 모절의 시제와 인칭이다. 시제에서는 현재를 기준그룹으로 하고 1은 과거, 2는 미래에 해당한다. 인칭은 3인칭을 기준그룹으로 1은 1인칭 2은 2인칭 3은 알 수 없음에 해당한다. 표의 각 열에 대해서는 순서대로 회귀계수 추정치, 표준오차, Z값, p-value, 오즈비, 오즈비의 2.5%신뢰구간, 97.5%신뢰구간을 의미한다.

내포절의 시제와 인칭이 축소모델에 반영되지 않은 것은 해당 독립변수들이 함의 관계를 설명해내는데 미치는 영향력이 적었기 때문이다. <표6 >을 통해 알 수 있는 것은 조건, 부정, 의문 유형의 함의 취소유형소 그리고 모절의 시제가 과거일 때, 모절의 인칭이 1인칭과 알 수 없을 때가 p-value가 유의 수준인 0.05보다 낮아 함의 관계에 영향을 줄 수 있는 요인이 될 가능성이 있다는 것이다. 회귀 계수와 오즈비는 추후의 분석과정에서 해석하도록 할 것이다.

4.2.3 모형 적합도 검정

모델 선택 과정에서는 독립변수들의 조합에 따른 여러 모델들 중 AIC를 비교하였다. 해당 절에서는 모형이 자체로 데이터를 잘 설명하고 있는지를 Hosmer-Lemeshow 검정을 통해 확인한다. Hosmer-Lemeshow 검정은 로지스틱 회귀 모델의 적합도를 평가하는 데 사용되는 통계적 검정 방법이다. 이 검정은 모델이 데이터에 잘 적합되었는지를 확인하며, 예측된 확률

	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)

표 6 변수 선택 후의 축소모델의 로지스틱 회귀 분석 결과

과 실제 관측된 결과 간의 차이를 평가한다. Hosmer-Lemeshow 검정은 다음과 같은 절차를 통해 모델의 적합도를 평가한다. 먼저, 관측치를 예측된 확률에 따라 여러 그룹으로 나눈다. 각 그룹 내에서 예측된 확률과 실제 발생한 사건의 비율을 비교하여 모델의 예측력을 평가한다. 마지막으로 각 그룹에서의 차이를 결합하여 검정 통계량을 계산하고, 이를 통해 모델의 적합도를 평가한다.

Hosmer-Lemeshow 검정에서 추정된 로지스틱 모형이 적합하면 근사적으로 카이제곱 분포를 따르게 된다. 검정의 귀무가설이 '추정된 모형이 잘 적합하다.'이기 때문에 모형이 적합하려면 귀무가설을 채택해야한다. 카이제곱 검정통계량이 작고 p-value가 크면 귀무가설을 기각할 수 없다. 축소모형의 Hosmer-Lemeshow 검정 결과 카이제곱 통계량은 7.79이고 p-value가 0.09258이므로 유의 수준 (0.05)보다 크므로 귀무가설을 기각할 수 없다. 즉 모형이 함의 관계 설명에 잘 적합되었다고 할 수 있다.

X^2	Degree of freedom	p-value
7.7927	4	0.09258

표 7 Hosmer-Lemeshow 검정 결과

4.2.4 독립변수의 유의성 검정

각 독립변수가 유의한지를 검정하기 위해 Wald 검정을 진행하였다. Wald 검정은 로지스틱 회귀 모델에서 독립변수의 유의성을 평가하는 검정법으로 왈드 통계량(W)를 사용한다. 왈드 통계량을 다음과 같이 정의되고 $W=(\frac{\beta-\beta_0}{SE(\hat{\beta})})^2$ 특정 회귀계수 추정값을 표준 오차로 나눈 값의 제곱이다. 주로 특정 회귀계수가 0이라는 귀무가설을 검정하는 데 사용된다. 따라서 p-value가 유의 수준(0.05)보다 작다면 귀무가설을 기각할 수 있고, 이는 회귀계수가 0이 아님을 즉 해당 독립변수가 통계적으로 유의미함을 의미한다. 검정결과 함의 취소 운용소 유형과 모절의 시제 및 인칭이 유의한 변수로 확인되었다. 축소모델에서 사용된 독립변수들과 모저과 내포절의 주어 일치여부를 제외하고는 일치된 모습을 보인다. 이러한 경우에는 모절과 내포절의 주어 일치 여부와 다른 변수간의 상호작용이 있으리라 추측해볼 수 있다. 두 변수간의 상호작용이 있을 경우, 개별 변수는 유의하지 않지만 상호작용 항이 유의할 수 있다. 이 경우 개별 변수는 모델에서 중요한 역할을 하기 때문에 포함될 수 있다. 더불어 (Walsh & John)의 선행연구를 고려한다면, 해당 변수는 함의와 모순을 분류해내는데 기여하기 보단 함의 관계 자체에 영향을 주기에 축소모델에는 포함되었지만 개별 변수는 유의하지 않은 것일 수 있다. 만약 내포절과 모절의 주어가 일치한다면 함의 관계가 모순인지 함의인지에 관계 없이 함의 관계를 파악하는 정확도와 속도에 영향을 준다는 것이다.

4.2.5 오즈비(Odds Ratio) 비교

함의 취소 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

표 8 Wald 검정 결과

오즈비는 두 사건의 오즈(odds)를 비교한 비율이다. 오즈는 어떤 사건이 발생할 확률과 발생하지 않을 확률의 비율로 정의된다. 오즈비는 한 변수에 대해 처치를 가했을 때의 사건의 발생 확률과 가하지 않았을 때의 사건의 발생확률에 대한 오즈의 비율로 정의된다.

$$Odds = \log\left(\frac{P(Y=1)}{P(Y=0)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$$

$$Odds\ Ratio\ (OR) = \log\left(\frac{P(Y=1|X=1)}{P(Y=0|X=1)} - \frac{P(Y=1|X=0)}{P(Y=0|X=0)}\right)$$

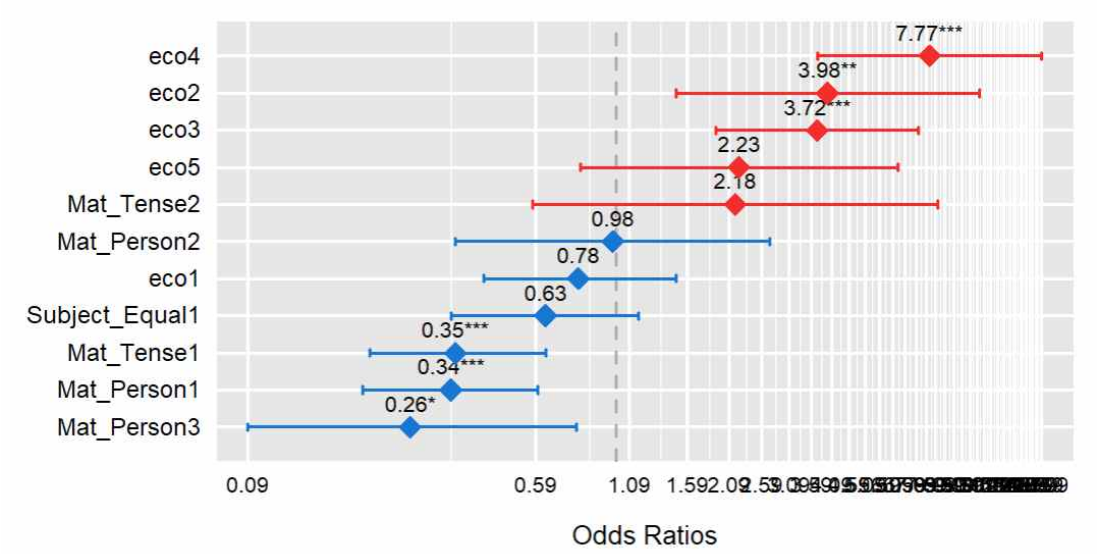

그림 1 축소모형의 오즈비. Y축의 변수명은 순서대로 부정, 의문, 조건, 2개 이상, 모절 미래시제, 모절 주어 2인칭, 양태, 모절과 내포저의 주어 일치, 모절 과거시제, 모절 주어 1인칭, 모절 주어 알 수 없음을 의미한다.

축소모형에 이용된 독립변수들의 오즈비는 <그림 1>에 나타나 있다. 그래프에서 별표는 변수의 유의성을 의미한다. 별표가 1개있으면 유의수준 0.05수준에서 해당 변수가 유의함을 별표가 3개 있으면 유의수준 0.001에서 해당 변수가 유의함을 의미한다. 따라서 오즈비에 관계없이 함의 취소 운용소의 유형이 양태, 2개 이상인 경우 그리고 모절의 시제가 미래인 경우와 인칭이 2인칭인 경우 그리고 모절과 내포절의 주어가 일치하는 경우(eco1, eco5, Mat_tense2, Mat_Person2, Subject_Equal)는 함의 관계를 설명하는데 유의하지 않으므로 이후 설명에서 배제할 것이다.

함의 취소 운용소 유형이 조건, 의문, 부정인 경우의 오즈비는 1보다 크며 95% 신뢰구간도 모두 1을 넘는다. 이는 변수들이 모순의 가능성을 높이는 긍정적인 영향을 미친다는 것을 의미한다. 함의 취소 운용소 유형이 부정인 경우의 오즈비는 약 7.77로 이는 해당 변수 값이 1 증가할 때 종속변수가 1인 사건이 발생할 가능성이 7.77배 높다고 해석할 수 있다.. 모든 독립변수는 더미변수로 함의 취소 운용소 유형이 부정에 해당하면 1, 아니면 0을 갖도록 재코딩되었고 종속변수는 모순일 때 1의 값을 가진다. 그리고 기준집단은 해당 범주에 대해 모든 항목이 0이다. 따라서 함의 취소 운용소 유형이 부정일 때 모순으로 예측될 가능성이 기준집단인 함의 취소 운용소가 없는 경우에 비해 7.77배 높음을 의미한다. 의문과 조건인 경우도 마찬가지로 각각 모순으로 예측될 가능성이 3.98배 3.72배 높아진다.

모절의 시제가 과거인 경우와 인칭이 1인칭과 알 수 없는 경우의 오즈비는 1보다 작으며 신뢰구간도 1미만이다. 이는 이 변수들이 함의의 가능성을 높이는 긍정적인 영향을 미친다는 것을 의미한다. 모절의 시제가 과거인 경우의 오즈비는 약 0.35로해당 변수 값이 1 증가할 때 종속변수가 1인 사건이 발생할 가능성이 약 6배 감소한다고 해석할 수 있다. 여기서 0.35배가 아니라 약 6배인 이유는 오즈비가 로그 스케일이기 때문이다. 로그스케일인 오즈비에서 0.1과 10이 종속변수에 영향을 미치는 정도는 비슷하다. 따라서 함의 취소 운용소 유형이 부정일 때 함의로 예측될 가능성이 기준집단인 모절의 시제가 현재인 경우에 비해 6배 높음을 의미한다. 모절의 인칭이 일인칭과 알 수 없는 경우도 마찬가지로 각각 모순으로 예측될 가능성이 약 6배 약 7배 높아진다.

4.3 로지스틱 회귀분석 결과 요약 및 가설 설정

모델 선택과 적합성 검정을 통해 함의 관계에 기여하는 변수들의 최적 결함은 함의 취소 운용소 유형, 모절의 시제 및 인칭, 모절과 내포절의 주어 일치 여부임이 확인되었다.

Wald검정 결과 모절과 내포절의 주어 일치 여부는 개별적으로는 함의 관계를 설명하는데 있어 유의하지 않은 것으로 확인되었다. 하지만 축소모델에 반영됨으로써 모순과 함의로 분류하는 것이 아닌 함의 관계를 파악하는 것 자체에는 기여할 수 있는 가능성이 있다.

오즈비 분석 결과 함의 취소 운용소 유형이 조건, 의문, 부정인 경우는 함의 관계가 모순이 될 가능성을 높였으며, 모절의 시제가 과거인 경우와 인칭이 1인칭과 알 수 없는 경우는 함의 관계가 함의가 될 가능성을 높였다.

로지스틱 회귀분석 결과에 따라 세울 수 있는 국어 내포문의 함의 관계에 대한 가설은 다음과 같다.

> 가) 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.
>
> 나) 확신성 담화에 있어 함의 취소 운용소 유형이 없고, 모절의 시제가 과거, 인칭이 1인칭 또는 알 수 없는 경우에 확신성 담화와 내포 명제가 함의 관계를 가질가능성이 높다.

5. 언어모델을 통한 가설 검정

로지스틱 회귀 분석 결과에 따라 제시한 가설에 따라 실제 함의 관계가 변화하는지를 확인하기 위한 과정으로 언어모델을 사용할 것이다.

5.1 언어모델과 파인 튜닝 과정 소개

가설 검정에 사용할 언어모델은 KcELECTRA을 사용한다. KcELECTRA는 기존의 한국어 트랜스포머 기반 모델이 한국어 위키, 뉴스, 기사 책 등 잘 정제된 데이터를 기반으로 학습된 모델이기에 신조어, 오탈자 등 공식적인 글쓰기에 나타나지 않는 표현들에 대해서 잘 파악하지 못한다는 한계를 보완하기 위해 네이버 뉴스의 댓글과 대댓글을 수집해 토크나이저와 ELECTRA모델을 처음부터 학습한 Pretrained ELECTRA 모델이다. '추론_확신 분석 말뭉치'는 뉴스와 같은 문어 뿐 아니라 구어에서 수집한 데이터도 많이 포함하고 있기에 KcELECTRA를 선정하였다.

파인 튜닝을 위한 데이터는 국립국어원 모두의 말뭉치 인공지능(AI) 말평에서 제공하는 함의 분석 평가용 말뭉치를 사용한다. 해당 코퍼스는 언어 주석 정보가 달려있지 않아 로지스틱 회귀분석의 데이터로 사용하기엔 부적절 하지만 확신성 담화와 내포명제 그리고 함의 관계에 대한 정보 13521개로 구성되어 있으므로 모델을 Senstence Pair Classification 태스크를 수행하도록 파인 튜닝 시키기엔 적절하다. 해당데이터를 8:2로 나누어 훈련데이터와 테스트 데이터를 구성하였다.

파인 튜닝 후 2750개로 구성된 테스트 데이터에 대한 정확도와 F1스코어는 각각 0.692와 0.605이다. 정확도는 테스트 데이터 중 올바르게 예측된 것들의 비율로 0.692는 테스트 데이터의 약 69.92%가 함의 또는 모순으로 올바르게 분류되었음을 말한다. 동일한 데이터 셋에 대해 구글의 Multilingual BERT (M-BERT)가 0.48, KoBERT가 0.46, KR-BERT가 0.47[3]의 정확도를 갖는 것[4]을 고려하면 모델이 꽤 정확하게 분류해낸 다고 볼 수 있다. F1은 Recall (실제 TRUE인 것 중에서 모델이 positive라 맞게 예측한 것의 비율)과 Precisicion(모델이 positivie라 예측한 것 중 실제로 TRUE인 것의 비율)의 조화평균으로 1에 가까울 수록 모델 성능이 좋음을 말한다. 0.605는 바교적 높은 값으로 모델이 예측을 잘 했을 뿐아니라 함의와 모순에 치우치지 않고 균형적으로 분류했음을 의미한다.

5.2 가설에 따른 데이터 수정 및 언어모델의 예측 결과

코퍼스에서 선정한 몇 개의 데이터에 대해 가설에 따라 수정을 진행하였다.

가설 (가)를 검증하기 위해서 함의 취소 운용소가 없거나 양태인 경우 그리고 모절의 시제가 과거나 미래인 경우 그리고 인칭이 1인칭, 2인칭 그리고 알 수 없는 경우 중 함의로 분류된 데이터에 대해 함의 취소 운용소 중 부정, 의문, 조건 중 하나를 갖게 하고 모절의 시제를 현재로, 인칭을 3인칭으로 변경하였다. 수정된 데이터의 예시는 <표 9>과 같다.

---

3) 송상헌 외. 2021 말뭉치 함의 분석 및 연구. 국립국어원. 2021. 81-83

함의 취소 운용소 유형	기존의 확신성 담화	수정된 확신성 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 **있겠습니다.** 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 나타나 있었고요. 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 나타나 있었고요. 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 거를 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 거를 이제 **인식하지 않는 상황입니다.**

표 9 가설 (가) 검정을 위한 수정 데이터 예시

수정 후 함의 취소 운용소가 의문, 조건, 부정인 데이터는 각각 8, 9, 9개로 문장의 자연스러움을 고려하여 배정하였다. 총 26개의 수정 데이터 중 언어모델은 2개를 제외한 24개를 모순으로 분류하였다. 이는 약 92%가 함의에서 모순으로 변화한 것이며, 이에 따라 가설 (가)를 수용할 수 있을 것이다.

가설 (나)를 검증하기 위해서 함의 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 데이터에 대해 함의 취소 운용소를 없애고 모절의 시제를 과거로 인칭을 1인칭

또는 알 수 없도록 변경하였다. 수정된 데이터의 예시는 <표 10>과 같다.

수정 후 모절의 인칭이 1인칭, 알 수 없음인 데이터는 각각 11, 15개로 문장의 자연스러움을 고려하여 배정하였다. 총 26개의 수정 데이터 중 언어모델은 6개만을 함의로 분류하였다. 이는 약 24%가 모순에서 함의로 변화한 것이다. 따라서 가설 (나)는 수용하기에 어려움이 있다.

가설 (나)에 대한 수정 데이터에서는 가설이 유효하지 않은 이유는 두가지를 고려할 수 있다. 첫째는 가설이 함의 관계를 잘 설명하고 있지 않아서, 둘째는 언어모델이 인간의 언어 사용을 잘 반영하고 있지 않아서이다. 차후에 추가적인 데이터로 2차검증을 시도하거나 언어모델이 아닌 실제 국어 화자를 대상으로 실험을 한다면 가설이 정말 유효하지 않은지를 다시 확인할 수 있을 것이다.

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. 이때 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. 이때 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.**

표 10 가설 (나) 검정을 위한 수정 데이터 예시

6. 결론

'추론_확신 분석 말뭉치'에 대해 함의 관계를 종속 변수로 6가지의 언어학적 정보를 로지스틱 회귀분석을 이용하여 분석한 결과 함의 관계에 기여하는 변수들의 최적 결합은 함의 취소 운용소 유형, 모절의 시제 및 인칭, 모절과 내포절의 주어 일치 여부임이 확인하였고 오즈비 분석을 통해 함의 취소 운용소 유형이 조건, 의문, 부정인 경우는 함의 관계가 모순이 될 가능성이 높아짐을 모절의 시제가 과거인 경우와 인칭이 1인칭과 알 수 없는 경우는 함의 관계가 함의가 될 가능성이 높아짐을 확인했다.

이를 바탕으로 가설을 세워 Sentence Pair Classification 태스크 수행을 위해 내포문과 그

에 따른 함의 관계로 파인 튜닝한 한국어 언어모델에 대해 검정을 진행하였다. 그 결과 모순으로 분류될 가능성을 높이는 조합에서는 가설이 유의하였지만, 함의로 분류된 가능성을 높이는 조합에서는 가설이 유의하지 않음이 확인되었다.

한편 본 연구는 코퍼스에 포함된 언어 정보로만 함의관계를 설명하고 있다는 한계가 있다. 이 연구에서 독립변수로 설정한 모절과 내포절의 주어 일치 여부, 함의 취소 운용소 유형, 내포절과 모절의 인칭과 시제 뿐 아니라 함의 관계에 영향을 미칠 수 있는 여러 요소들을 더 고려해볼 수 있다. 이를테면 술어의 사실성 여부 혹은 일반성이나 보편성을 표현하는 총칭적 표현의 포함 여부이다. 이들은 함의 관계에 영향을 줄 수도 있으나 말뭉치에 포함되지 않았기에 분석이 어려웠다. 또한 언어모델을 이용한 가설 검정을 수행함에 있어 언어 모델이 인간의 언어 사용을 잘 반영하고 있다는 전제가 필요하다는 한계가 있다. 해당 주장은 여전히 논란의 소지가 있다. 마지막으로 가설 검정을 위한 데이터 수정과정에서 연구자 개인의 직관에 호소해서만 수정을 진행하였으므로 신뢰성에 문제가 생길 수 있다.

위와 같은 한계를 보완하기 위해서 추후에 언어 정보 주석 추가, 언어모델이 아닌 언어 사용자에게 수정 데이터에 대한 확신성 판단 실험 수행, 수정 데이터 교차검증을 고려해 볼 수 있다. 본 연구는 코퍼스 데이터와 통계적 방법을 통해 언어 사용의 경향성을 파악하는데에 그쳤으므로 해당 현상이 발생한 원인에 대한 이론적 논의로 확장할 수 있을 것이다.

참고 문헌

- 김소정 외. 말뭉치 함의 분석 및 연구. 국립국어원. 2020
- 송상헌 외. 2021 말뭉치 함의 분석 및 연구. 국립국어원. 2021
- Walsh CR, Johnson-Laird PN. Co-reference and reasoning. Mem Cognit. 2004 Jan;32(1):96-106.
- de Marneffe, M.-C., Simons, M. & Tonhauser, J. (2019), The CommitmentBank: Investigating projection in naturally occurring discourse. Proceedings of Sinn und Bedeutung 23, 107-124.
= Kiparsky K. & P. Kiparsky(1971), Fact In D. Steinberg and L. Jakobovits(eds.) Semantics: An Interdisciplinary Reader in Philosophy Linguistics and Psychology, Cambridge University Press, 345-69
- “beomi/KcELECTRA-base”, Hugging Face, 2024년 6월 23일 접속, https://huggingface.co/beomi/KcELECTRA-base