단서 구문과 어휘 쌍 확률을 이용한 인과관계 추출*

한국과학기술원 전산학과,[1] 전문용어언어공학연구센터,[2] KT 기술연구소[3]
장 두 성[1,3] · 최 기 선[2†]

Causal Relation Extraction Using Cue Phrases and Lexical Pair Probabilities

Du-Seong Chang, Key-Sun Choi
Department of Computer Science, KAIST[1]/KORTERM[2]/KT[3], Deajeon, Korea

요 약

현재의 질의응답 시스템은 TREC(Text Retrieval Conference) 질의집합에 대해 최대 80% 정도의 응답 성공률을 보이고 있다.[1)] 하지만, 질의 유형에 따라 성능의 많은 차이가 있으며, 인과관계에 대한 질의에 대해서는 매우 낮은 응답 성공률을 보이고 있다. 본 연구는 인접한 두 문장 혹은 두 명사구 사이에 존재하는 인과관계를 추출하고자 한다. 기존의 명사구 간 인과관계 추출 연구에서는 인과관계 단서구문과 두 명사구의 의미를 주요한 정보로 사용하였으나,[2,3)] 사전 미등록어가 사용되었을 때 올바른 선택을 하기 어려웠다. 또한, 학습 코퍼스에 대한 인과관계 부착과정이 선행되어야 하여,[3)] 다량의 학습자료를 사용하기가 어려웠다. 본 연구에서는 인과관계 명사구 쌍에서 추출된 어휘 쌍을 기존의 단서구문과 같이 사용하는 방법을 제안한다. 인과관계 분류를 위해 나이브 베이즈 분류기를 사용하였으며, 비지도식 학습과정을 사용하였다. 제안된 분류 모델은 기존의 분류 모델과 달리 사전 미등록어에 의한 성능 저하가 없으며, 학습 코퍼스의 인과관계 분류 작업이 선행될 필요 없다. 문장 내 명사구 간의 인과관계 추출 실험 결과 79.07%의 정확도를 얻었다. 이러한 결과는 단서구문과 명사구 의미를 이용한 방법에 비해 6.32% 향상된 결과이며, 지도식 학습방식을 통해 얻은 방법과 유사한 결과이다. 또한 제안된 학습 및 분류 모델은 문장간의 인과관계 추출에도 적용 가능하며, 한국어에서 인접한 두 문장간의 인과관계 추출 실험에서 74.68%의 정확도를 보였다.

서 론

인과관계(Causality)는 일반적으로 어떤 사실과 다른 사실 사이의 원인과 결과관계를 말한다. 질의응답의 많은 수가 사건 간의 인과관계를 묻는 질문이며, 현재의 질의응답 시스템에서는 이러한 인과관계 질의에 대해 낮은 응답 성공률을 보이고 있다.

TREC에서 질의응답의 성능 평가를 위해 사용되는 질의는 사실에 근거한 짧은 답변을 요구한다. 최근의 질의응답 연구결과는 이러한 질의의 80% 정도에 대해 정답을 5개 후보 내에 제시할 수 있는 성능을 보이고 있다.[1)] 하지만, 질의응답 성능은 질문의 종류에 따라 많은 성능차를 보이고 있다.

(Moldovan et al., 2003)에 따르면, '누가', '언제', '어디서', '무엇' 등을 묻는 질의에 대해서는 상대적으로 높은 성능을 보이고 있다. 반면 '왜'와 같이 인과관계를 묻는 질의에 대해서는 매우 낮은 성능을 보이고 있다.[4)] 인과관계를 묻는 질의의 예는 다음과 같다.[1]

(TREC-902) "Why does the moon turn orange?"

(TREC-1103) "What is the effect of acid rain?"

(TREC-1141) "What is the effect of volcanoes on the climate?"

(TREC-1745) "What are hiccups caused by?"

*본 연구는 한국과학기술기획평가원 국책연구개발사업(M1-0107-00-0018)과 한국과학재단 특성화장려연구사업(R21-2003-000-10042-0)의 지원으로 수행되었습니다.

†E-mail : dschang@world.kaist.ac.kr

1 예문은 질의응답 성능평가 학회인 TREC-10/11에 실제 사용된 질의에서 발췌하였으며, 예문의 번호는 TREC의 질의 고유 번호이다.

TREC-8/9/10의 인과관계 질의에 대한 응답 성능은 0.031 MRR(Mean Reciprocal Rank)[2]로 매우 낮은 값을 보였다. 인과관계 처리에 한계가 있음에도 불구하고 전체 성능이 비교적 높은 이유는 사실에 근거한 단답형 질문 위주로 평가해 왔기 때문으로 해석된다.

사용자 간에 질의 응답을 주고 받는 인터넷 사이트[3]에 등록된 95만 질의 DB 중 13만개가 인과관계를 묻는 질의로 전체에서 많은 수를 차지하고 있다. 2003년의 TREC에서는 이러한 실제 질의응답의 요구를 반영하여 주어진 문서에서 인과관계나 방법에 대한 질문을 강조하고 있다. 또한, 2005년까지의 질의응답 로드맵에 따르면 전문지식 간의 관계를 파악하고 추론을 통해 새로운 지식을 추출하여 응답할 것을 요구하고 있다.[5)]

"What are hiccups caused by?"과 같이 인과관계를 묻는 질의에 응답하기 위해서는 몇 가지 문제들을 풀어야 한다. 그 첫째는 **'사건 추출'**로서, 질의에 표현된 결과 사건 'hiccups'를 포함하는 문장에서 사건들을 추출한다. 그 둘째는 **'인과관계 분석'**으로서, 추출된 사건 간의 인과관계를 분석한다. 마지막으로 **'인과관계 질의응답'**으로서, 문헌에서 분석된 인과관계에서 추론을 통해 주어진 질의에 대한 답변을 생성한다. 본 논문에서는 사건 추출과 인과관계 분석 문제에 중점을 두고자 한다.

두 사건간의 인과관계는 사건들을 연결하는 단서 구문을 이용하여 추정할 수 있다. 명사구로 표현된 사건 간 인과관계 단서구문은 명사구를 구문적으로 연결하는 동사구이다.[2,3)] 문장으로 표현된 사건 간 단서구문은 두 문장을 연결하는 문장 접속 부사나 한국어에서의 연결어미 등이 있다.

원인과 결과 사건 쌍으로부터 인과관계를 이끄는 어휘 쌍을 추출할 수 있으며, 이들 어휘 쌍의 확률을 같이 사용하여 단서 구문만을 사용하였을 때보다 인과관계 추출의 정확도를 높일 수 있다. 어휘 쌍 확률은 코퍼스로부터 학습할 수 있다.

2장에서는 인과관계 추출을 위한 기존의 연구들에 대해 살펴보고, 3장에서 인과관계 추출을 위한 모델을 제시한다. 4장에서는 문장 내 명사구 간 인과관계 추출 실험을 통해 제안된 모델의 성능을 분석하고 5장에서 제안된 모델이 문장 간 인과관계 추출을 위해서도 사용될 수 있음을 실험을 통해 보인다.

2 $NRR=\frac{\sum\frac{1}{R_i}}{N}$
N=질의의 수, Ri=i번째 질의에 대한 정답이 나타난 정답후보의 순위[6)]

3 네이버 지식iN, http://kin.naver.com

관련연구

1. 인과관계

문헌 내에서 인과관계는 다양한 형태로 표현된다. 인과관계는 아래 예문 (1a)과 같이 하나의 문장 내에서 주어와 목적어 간에 원인과 결과의 형태로 표현되기도 하며, 예문 (1b), (1c)과 같이 두개의 문장 혹은 절 간에 나타나기도 한다. 또한, (1d)와 같이 하나의 명사구 내에서 두 명사 간에 나타나기도 한다.[4]

(1a) "Earthquakes generate tidal waves."

(1b) "The meaning of a word can vary a great deal depending on the context. For this reason, pocket dictionaries have a very limited use."

(1c) "The traffic was so heavy that I couldn't arrive on time."

(1d) "disease-causing bacteria"

위 예문들에서 인과관계는 각각 동사(generate), 문장접속구문(for this reason, that), 복합명사 등을 이용하여 표현되었다. 본 연구에서는 예문 (1a)와 같이 하나의 문장 내에서 명사구간에 존재하는 인과관계의 추출을 첫 번째 목적으로 하며, 예문 (1b), (1c)와 같이 두 문장 간에 존재하는 인과관계의 추출에도 제안된 모델이 사용될 수 있음을 보인다.

2. 문장 간 인과관계 분석 연구

문장 간에 존재하는 인과관계는 수사구조의 일부로서 해석이 가능하다. (Marcu et.al., 2002)에서는 문장 간의 어휘 쌍 확률을 이용하여 57%의 문장 간 인과관계 분석 정확도를 보였다.[7)] 훈련에 사용된 문장들은 'Because of', 'Thus'의 단서 구문으로 연결된 문장들이며, 두 문장에 포함된 주요한 단어들[5]의 쌍에 대해 어휘 쌍 확률을 추출하였다. 여기에서 사용한 어휘 쌍 확률은 문장 내 명사구 간에서도 사용이 가능하리라 추정된다.

3. 명사구 간 인과관계 분석 연구

(Girju et al., 2002, 2003)의 인과관계 추출 연구에서는 명사구들의 의미부류와 인과관계를 이끄는 동사구를 단서로 인과관계를 추출하여, 질의응답의 성능을 향상시키는데

4 각각의 사건을 기술하는 두 문단 사이에도 인과 관계는 존재할 수 있다. 이러한 관계는 수사 구조 해석의 문제로 본 논문에서는 논외로 한다. 예문 (1a)~(1d)는[2)]에서 발췌하였다.

5 어휘 쌍의 구성에 참여할 주요 단어로 명사, 동사, 형용사만을 사용하였다.

사용하였다. 문장 내 명사구 간의 인과관계를 이끄는 동사구는 WordNet으로부터 추출하였다. 인과관계의 사건으로 사용될 수 있는 명사 의미부류를 WordNet의 최상위 명사 개념에서 추출하여 이를 인과부류(Causation Class)[6]라 정의하였다. <명사구, 동사, 명사구>의 인과 관계 후보는 명사구들이 어떠한 의미부류에 속하는가에 따라 다섯 등급(이하, 명사 분류 값)으로 결정된다. 이중 4등급까지를 인과관계로 분류하는 실험에서는 65.5%의 정확도를 보여 주었으며,[2] 결정 트리를 이용한 지도식 학습 방법을 사용하여 73.91%의 정확도를 보였다.[3]

4. 토 론

인과관계 지도식 학습을 위해서는 코퍼스에 대한 인과관계 부착과정이 선행되어야 한다. 하지만 이러한 과정은 많은 시간을 소요하는 과정이므로 다량의 학습 자료를 사용하기 어려운 면이 있다. 본 연구에서는 인과관계가 부착되지 않은 코퍼스로부터 비지도식 학습을 통해 인과관계 어휘 쌍을 추출하여 이를 기존의 단서구문 및 명사 분류 값과 같이 사용하는 방법을 제안한다.

인과관계 분류의 근거로 WordNet이나 사전을 사용할 때의 단점은 사전에 등록되지 않은 단어에 대해 명사부류를 정의할 수 없다는 것이다. 실제 4장에서 사용된 실험 집합에서 44개의 미등록어가 발생하였으며 이중 36.4%인 16개로부터 오분석이 유도되었다.[7] 이 문제를 해결하기 위해서는 전문용어/고유명사 분류기를 같이 사용하는 방법이 있을 수 있다. 본 연구에서는 어휘 쌍 확률을 같이 사용하여 이러한 문제를 해결하였다.

명사구 간 인과관계 추출

인과관계의 추출 문제는 두 인과관계 후보 사건 쌍에 대하여 인과관계가 존재할 경우와 존재하지 않을 경우로 분류하는 문제로 해석할 수 있다. 인과관계 분류를 위해 나이브 베이즈 분류기를 사용하며, 단서구문과 어휘 쌍 확률을 분류를 위한 속성으로 사용하였다.

1. 인과관계 추출을 위해 사용 가능한 속성들

아래 예문 (3a)~(3c)은 문장 내 명사구 간 인과관계가 존재하는 문장이다.[8]

(3a) Unprotected sun **exposure** will cause premature **aging** of the skin.

(3b) Skin **cancer** usually appears in adulthood, but it is caused by sun **exposure** and **sunburns** that began in childhood.

(3c) Long-standing mouth **ulcers** may develop into oral **cancer**.

명사구 간 인과관계는 몇 가지 단서로서 추정이 가능하다. 첫번째 단서로 두 명사구를 구문적으로 연결하는 동사구의 의미이다. 위 예문에서 'cause', 'develop' 등의 동사구는 "A가 B를 야기하다."와 같은 의미를 가지고 있다. 단서로서 사용될 수 있는 두 번째는 두 명사구에 사용되는 어휘 쌍이다. 'sun exposure'나 'sunburn'은 'skin cancer'의 원인이 된다고 알려져 있으므로, 이 어휘 쌍의 존재는 예문 (3b)가 인과관계를 기술하고 있다는 훌륭한 증거가 될 수 있다.[9] 마지막으로는 두 명사구가 속하는 개념 분류이다. 즉, 'ulcer'의 상위어인 'pathology'는 'cancer'의 상위어인 'unhealthiness'와의 의미관계로부터 하위 단어들의 원인-결과 관계를 예측할 수 있다.

본 논문에서는 <명사구, 동사, 명사구>의 인과관계 후보에서 인과관계를 추정하기 위해 동사구의 어휘와 두 명사구간에 존재하는 어휘 쌍의 확률을 사용한다.

2. 의존구조에서 인과관계 후보 주출

Fig. 1은 예문 (3b)의 의존 구조이다.

이와 같은 의존 구조에서 인과관계의 후보가 되는 <명사구, 동사, 명사구> 형태의 3진 관계를 추출하기 위해서는 명사구 추출, 동사 추출, 명사 참조 해결, 동격 명사구 탐색, 3진 관계 추출 등의 단계를 거친다.

명사구와 동사구는 주어진 문장의 의존구조에서 명사 및

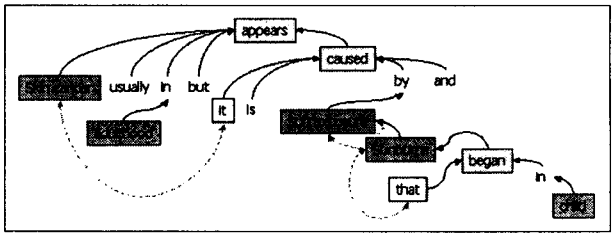

Fig. 1. 예문(3b)의 의존구조.

6 WordNet에서 [human action], [phenomenon], [event], [state], [psychological feature]로 표현된다.

7 그 결과, 실험집합(cMedline)에 대한 기준 시스템 성능에서 정확도와 재현율이 각각 15.15%, 11.10% 하락하였다. WordNet에는 고유명사 및 신조어가 포함되어 있지 않기 때문에 질병영역 문서와 같은 전문용어가 많이 사용된 문서를 대상으로 인과관계를 추출하고자 할 때 이러한 문제가 크게 나타난다.

8 예문은 Medline 질병백과사전에서 추출하였다.

9 이와 같은 어휘 쌍이 명사구내에 존재하더라도 인과관계가 성립하지 않는 경우도 있다.
예 : The National Cancer Institute reported the safety notice on sunburn.

동사가 지배소인 부분 트리를 탐색하여 추출한다. 탐색 된 명사구 중 관계 대명사의 경우 의존 구조에서 선행사를 탐색하여 이를 3진 관계 생성에 반영한다. 그 결과 'sun exposure and sunburn'과 'child'가 'begin in'으로 연결되는 3진 관계의 재료로 고려된다. 대명사가 참조하는 명사는 같은 구문 트리 내에서 (Sidner, 1981)의 초점이론[10]에 근거하여 찾았다. 동격 명사구의 탐색은 의존 구조에 표현되는 동격 구조를 참조하였다.

문장 내 사용된 동사로부터 주격 명사구와 목적격 명사구의 쌍을 찾아 3진 관계를 구성한다. 이 과정에서 자동사+전치사를 하나의 동사로 취급하였으며, 시간이나 장소를 나타내는 전치사구문은 여기에서 제외되었다. 결과적으로 Fig. 1의 의존구조에서 아래 2개의 3진 관계가 추출되었다.

(3d) < 'sun exposure', 'caused by', 'skin cancer' >

(3e) < 'sunburn', 'caused by', 'skin cancer' >

3. 인과관계 분류기

t_i가 문헌에서 추출된 3진 관계(인과관계 후보)라 할 때, t_i의 인과관계 부류 y_i가 c_j(∈{0, 1})일 확률은 수식 (1)에 의해 추정할 수 있다. f_{ik}는 3진 관계 t_i의 속성들로서 여기에서는 두 명사구간에 포함된 어휘 쌍들이다.

$$P(y_i = c_j \mid t_i;\hat{\theta}) = \frac{P(c_j \mid \hat{\theta})\prod_{k=1}^{|t_i|} P(f_{t_ik} \mid c_j;\hat{\theta})}{\sum_{r=0}^{1} P(c_r \mid \hat{\theta})\prod_{k=1}^{|t_i|} P(f_{t_ik} \mid c_r;\hat{\theta})} \quad (1)$$

파라미터 $P(c_j \mid \hat{\theta})$와 $P(f_{t_ik} \mid c_j;\hat{\theta})$는 인과관계의 여부 c_j가 표기된 대량의 3진 관계 리스트로부터 학습할 수 있다. 본 연구에서는 인과관계의 여부가 표시된 대량의 3진 관계 리스트 대신 인과관계의 여부가 검증되지 않은 초기 인과관계 쌍으로부터 기대치-최대화(Expectation-Maximization) 방법을 이용하여 비지도식 학습을 한다. 또한 학습 코퍼스에 나타나지 않는 어휘 쌍을 위해 Laplace 평활화(smoothing)를 적용하였다.

4. 인과관계 분류기 학습

인과관계 분류기의 학습 과정은 비지도식 학습과정을 사용한다. 학습단계는 3단계로 이루어지며, 1단계에서는 기존의 연구에서 제시된 단서구문과 명사 의미부류를 이용하여 초기 인과관계 명사구 쌍을 추출한다. 사용된 단서구문 및 명사 부류는(Girju et al., 2002)에서 제시된 바를 사용하였다. 여기에서 단서구문은 3진 관계가 인과관계로서 사용될 수 있는지를 판단하는 단서로 사용되는 동사구이다. 이 단서 구문은 WordNet에서 추출된 60개 단서구문으로 이루어져 있다.[11] 참조한 논문에서 인과관계 후보는 명사구 및 동사구가 인과관계에 사용될 수 있는지의 여부에 따라 다섯 등급으로 분류되었다. 초기 인과관계 쌍 추출을 위해서는 높은 신뢰도를 보장하기 위해 3등급까지만을 초기 인과 관계 쌍으로 추출하였다.

2단계에서는 추출된 인과 관계 쌍으로부터 어휘 쌍을 추출하여, 식 (1)에 사용된 파라미터를 추정한다. 3단계에서는 식 (1)과 1단계에서 사용된 단서구문 등을 같이 사용하여 보다 신뢰도 높은 인과관계 명사 쌍을 추출한다. 마지막으로 이 명사 쌍으로부터 다시 어휘 쌍을 추출하며, 주어진 학습 코퍼스에서 파라미터의 확률이 최대가 될 때까지 2, 3단계를 반복한다.

5. 인과관계 분류 모델

어휘 쌍 확률을 이용하여 학습된 분류기를 실제 적용할 때 단서구문 및 명사 분류 값과 같이 사용하는 방법에 따라 두 가지의 모델이 고려되었다.

분류 모델 BL+LP는 단서구문 및 명사 분류 값과 어휘 쌍 확률을 같이 사용하는 모델이다. 수식 (2)에서 $rank_i$는 2.3절에서 제시된 다섯 등급의 인과관계 분류이며, 명사 분류 값 $P(y_i|rank_i)$은 각 단계 별로{1.0, 0.8, 0.6, 0.4, 0.2}이 주어졌다. w는 가중치로 선험적으로 0.7이 주어졌다.

$$P_{BL+LP}(y_i \mid t_i) = (1-w)\times P(y_i \mid rank_i) + w \times P(y_i \mid t_i) \quad (2)$$

분류모델 CP+LP는 기본적으로 단서구문과 어휘 쌍 확률만 적용하는 방법으로, 단서구문과 어휘 쌍 확률로서 변별력 $Dist(t_i)$가 일정한 임계치 h 미만으로 떨어지는 경우에만 명사 분류 값을 사용한다.

$$P_{CP+LP}(y_i \mid t_i) = \begin{cases} P(y_i \mid t_i) & if\ Dist(t_i) \geq h \\ P(y_i \mid rank_i) & otherwise \end{cases} \quad (3)$$

명사구 간 인과관계 분류 실험

1. 실험 집합

인과관계의 훈련 및 실험을 위해 TREC에서 제공한 코퍼스의 일부를 사용하였다. 훈련에 사용된 코퍼스는 LA TIMES

10 대명사의 지시체가 될 수 있는 후보들을 초점 창고에 순서화 하여 보관하여 나가는 방법으로, 뒷 문장에서 대명사가 나왔을 때, 이 대명사의 지시체를 별도의 추론 없이 찾을 수 있도록 해 준다. 문장 내에 행위자(agent)로 사용된 대명사는 행위자 초점 리스트에서, 그 외의 대명사는 담화 초점 리스트에서 우선 찾는다.[8]

11 그 예는 다음과 같다.
'give rise to', 'induce', 'produce', 'generate', 'effect'

기사 187문단, Wall Street Journal 기사 203만 문단이다.

인과관계 분류의 성능 측정을 위해 두 가지 평가 집합을 사용하였다. 하나는 WSJ 1988년 기사 중 'cancer'를 포함하는 965문단(이후 cTREC으로 표기)이며, 질병영역 문헌인 A.D.A.M Inc.에서 제공하는 의학백과사전 중 'cancer'를 포함하는 518문단을(이후 cADAM으로 표기) 또 하나의 평가 집합으로 사용하였다.

2. 작업의 난이도

성능 평가용 코퍼스에 대해 두 명의 검증자[12]에 의해 각각 인과관계 여부를 표시하였다. 그 결과, 두 검증자 간에 의견이 일치한 경우가 72.59%였다. 이는 인과관계 추출 작업의 난이도를 나타낸다. 두 검증자 간에 의견이 다른 경우 협의를 통해 하나의 답안을 만들었다. Table 1은 검증 작업의 결과이다. Table에서 F-값은 모범답안을 기준으로 각 검증자의 분류결과에 대한 성능 평가이다.

검증 작업의 결과 높은 정확도에 비해 재현율이 매우 낮으며, 두 검증자간의 의견 일치가 상당히 힘들었음을 알 수 있다.

3. 훈련 및 실험 결과

Table 2는 3가지 방법에 따른 인과관계 분류 실험 결과이다. Table에서 BL은 기준 시스템으로서(Girju et al., 2002)에서 제시하는 단서구문과 명사 분류 만을 적용하여 분류한 결과이다. BL+LP는 단서구문 및 명사 분류 값과 어휘 쌍 확률을 같이 사용한 결과이며, CP+LP는 단서구문과 어휘 쌍 확률의 변별력이 떨어질 때만 명사 분류 값을 사용한 결과이다.

Table 1. 검증 집합에 대한 두 검증자의 작업 결과

cTREC	제시한 인과 관계 쌍의 수	올바른 인과 관계 쌍의 수	정확도	재현율	F-값
모범답안	142	142	100	100	100
검증자 1	116	114	98.27	88.37	89.06
검증자 2	92	92	100	71.31	78.63

Table 2. 명사구 간 인과관계 분류 실험 결과

분류모델	실험집합	정확도	재현율	F-값
BL	cMedline	66.92	42.75	53.08
	cTREC	83.64	65.71	73.60
	합계	74.37	54.61	62.98
BL+LP	cMedline	67.02	48.08	58.33
	cTREC	85.05	65.00	73.68
	합계	76.62	56.83	66.26
CP+LP	cMedline	72.73	61.07	68.97
	cTREC	85.71	64.29	73.47
	합계	79.07	62.73	69.96

12 한 사람은 본 논문의 첫 번째 저자이며, 다른 한 사람은 의학분야 전문가이다.

실험 결과, 단서 구문과 어휘 쌍 확률을 같이 사용한 경우 기준 시스템에 비해 정확도가 3.03% 상승하였으며, 어휘 쌍 확률이 변별력이 떨어질 때만 명사구 분류 값을 사용한 경우 정확도는 6.32% 상승하였다.

실험의 결과로 해석하였을 때, 단서구문의 여부는 인과관계를 검증하는 데 큰 도움이 되며, 비지도식 학습 방법에서 명사구 분류 값보다는 어휘 쌍 확률이 보다 인과관계 파악에 도움이 되고 있다는 것을 알 수 있다.

또한 2.4절에서 논의된 기준 시스템의 사전 미등록어에 의한 오류 중 37.5%가 제안된 분류모델(CP+LP)에서 교정되었다. 이는 제안된 모델이 사전 미등록어에 영향을 받지 않는 모델임을 보여준다.

제안된 분류 모델은 기존의 분류 모델에 비해 향상된 성능을 보이고 있으며, 비지도식 학습 방법으로 지도식 학습 방법을 통한 인과관계 분류기[3]와 유사한 성능을 보였다. 또한, 지도식 학습 방법과는 달리 분류기의 학습을 위해 다량의 인과관계 분류 코퍼스의 작성 과정이 필요 없다.

문장 간 인과관계 추출

두 사건 간의 인과관계는 두 사건을 연결하는 단서구문과 두 사건을 구성하는 어휘 쌍의 확률로서 추정이 가능하다. 앞 장에서는 영어 문장 내에서 명사구로 표현되는 사건들 간의 인과관계를 추정하였다. 이를 위해 두 명사구를 연결하는 동사구문을 단서 구문으로 사용하였다. 이 장에서는 제안된 모델을 한국어 문장 간에 존재하는 인과관계 추출을 위해 적용한다.

1. 인과관계 단서구문 추출

하나의 문장으로 표현되는 사건들 간에 존재하는 인과관계를 추정하기 위해서는 문장 접속 구문을 단서 구문으로 사용할 수 있다. 한국어에서 두 개의 문장을 연결하기 위해 사용되는 것은 문장 접속 부사, 문장 연결 어미 등이 있다.

(윤평현, 1989)에서는 문장간 연결어미로 연결되는 관계들 중 인과관계, 결과관계, 조건관계 등에 대해 다음과 같이 분류하였다. 인과관계의 종속절은 주절이 있게 한 객관적이고 일상적인 상태를 나타내며, 결과관계의 종속절은 주절의 시점에서 아직 일어나지 않은 내용으로 주절의 사건 수행 시 예상되어지는 결과를 나타낸다. 조건 관계의 종속절은 주절이 수행되기 위한 조건을 나타낸다.[9] 이들 관계

의 예는 아래 예문 (5a)~(5c)와 같다.

(5a) 하수구를 고쳐서 물이 잘 빠진다(인과관계).

(5b) 물이 잘 빠지도록 하수구를 고쳤다(결과관계).

(5c) 바람이 불면 낙엽이 떨어진다(조건관계).

본 연구에서는 이들 세 관계를 같이 인과관계라고 통칭하고, 이들 연결어미를 인과관계 단서구문의 기초로 삼았다. Table 3은 문장 간 인과관계 추출을 위한 단서구문의 일부이다. 전체 단서구문은 54개로 구성되어 있으며, 연결어미, 문장 접속 조사와 코퍼스에서 추출한 일부 문장 패턴으로 구성되어 있다. Table에서 <CAU>, <RSL>은 각각 원인 사건과 결과 사건이 놓일 위치를 표현한다.

단서구문의 신뢰도는 수식 (4)와 같이 정의된다.

$$P(y_{t_i} \mid CP_{t_i}) = P(\{ec_{t_i}, er_{t_i}\} \in CR \mid CP_{t_i}; D_{CR}) \quad (4)$$

단서구문 검증용 문헌 D_{CR}에 포함된 인과관계 후보 t_i가 <ec_{ti}, CP_{ti}, er_{ti}>의 3진 관계로 표현될 때, 단서구문의 신뢰도는 그 단서구문으로 추출된 문장 쌍 {ec_{ti}, er_{ti}}이 미리 분석된 인과관계 쌍(CR)에 포함될 확률이다. 여기에서, *ec*는 원인 문장, *er*은 결과 문장, *CP*는 단서구문이다. 단서구문 검증용 문헌은 코퍼스에서 각 단서구문 별로 5~20개 씩, 총 970개 문장 쌍으로 구성하였으며, 이들 문장 쌍에 대한 인과관계 여부는 수동 분석되었다.

2. 인과관계 후보 주출

문장 간 인과관계 후보는 단서구문에 의해 추출된다. 인과관계 후보는 <원인문장, 단서구문, 결과문장>의 형태로서, 예문 (5a)~(5c)로부터 추출할 수 있는 인과관계 후보는 다음과 같다.

(5d) < "하수구를 고치다", -어서, "물이 잘 빠진다.">

(5e) < "하수구를 고치다", -도록, "물이 잘 빠진다.">

(5f) < "바람이 불다", -면, "낙엽이 떨어진다.">

Table 3. 사용된 단서 구문의 예

신뢰도	단서 구문
1.0	<RSL>-는 이유는 <CAU>-기 때문이-
1.0	<RSL>-는데, 원인은 <CAU>-기 때문이-
0.9	<RSL> 이는 <CAU>-기 때문이-
0.85	<CAU>-는 경우에는 <RSL>
0.8	<CAU> 그래서 <RSL>
0.8	<CAU>-므로/니까/느라고 <RSL>
0.75	<RSL>-도록 <CAU>
0.7	<CAU>-려면/라면/다면/거든 <RSL>
0.6	<CAU>-면/니/기에/기 때문에 <RSL>
0.5	<CAU>-니까 <RSL>
0.4	<CAU>-아 <RSL>
0.2	<CAU>-경우 <RSL>
0.1	<CAU>-ㄹ 때 <RSL>
0.1	<CAU>-어서/아서 <RSL>

3. 인과관계 분류 모델

문장간 인과관계 분류를 위해서도 3장에 기술된 인과관계 분류기를 사용한다. 이 분류기의 비지도식 학습을 위해 수식 (4)의 단서구문의 신뢰도만을 이용하여 초기 인과관계 쌍을 추출한다. 초기 인과관계 쌍은 높은 신뢰도를 가지는 단서구문만을 이용하여 추출한다.[13]

사용된 인과관계 분류 모델은 수식 (5)와 같이 단서구문 신뢰도와 문장 쌍에 존재하는 어휘 쌍 확률을 사용한다. 식에서 $P(y_i \mid CP_{t_i})$는 인과관계 후보 t_i에서 사용된 단서구문 CP_{t_i}의 신뢰도이다.

$$P_{BL+LP}(y_i \mid t_i) = (1-w) \times P(y_i \mid CP_{t_i}) + w \times P(y_i \mid t_i) \quad (5)$$

4. 인과관계 분류기

인과관계 분류기의 구성은 Fig. 2와 같다. 코퍼스에서 구문분석을 통해 사건으로 사용될 수 있는 문장 혹은 명사구를 추출한다. 구문 트리에서 인접한 두개의 사건은 인과관계 사건 쌍의 후보로 추출되며, 3진 관계로 표현된다. 인과관계 분류기는 단서구문 및 어휘 쌍 확률을 이용하여 인과관계 분류를 하며, 분류된 사건 쌍은 다시 어휘 쌍 확률을 학습하기 위해 사용된다.[14]

5. 실 험

분류기 훈련을 위해 웹에서 제공되는 질병 백과 사전의 2158개 문서를 사용하였다. 훈련 문서에서 추출된 인과관

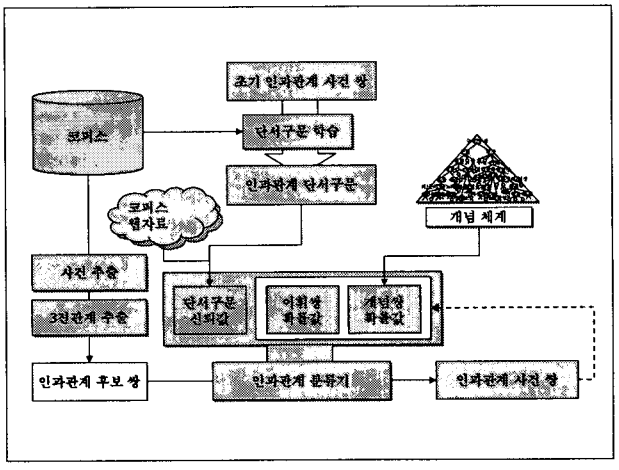

Fig. 2. 인과관계 분류기의 구성.

13 초기 인과관계 쌍의 추출을 위해 0.7이상의 신뢰도를 가지는 단서구문만을 사용하였다.

14 본 논문에서는 개념 쌍 확률 값은 실제 사용되지 않았다.

Table 4. 문장 간 인과관계 분류 실험 결과

분류모델	정확도	재현율	F-값
BL	68.57	52.17	59.25
BL+LP	74.67	64.13	69.01

계 후보 쌍은 3만여개이다. 분류기의 성능 평가를 위해 질병백과 사전에서 훈련세트에 포함되지 않은 4개의 문서를 사용하였다. 성능 평가용 문서에서 추출된 인과관계 후보 쌍은 122개이다.

Table 4는 2가지 방법에 따른 인과관계 분류 실험 결과이다. 표에서 BL은 기준 시스템으로서 단서구문의 신뢰도만을 사용하여 분류한 결과이다. BL+LP는 단서구문 신뢰도와 어휘 쌍 확률을 같이 사용한 결과이다.

실험 결과, 어휘 쌍 확률을 같이 사용했을 때 단서 구문만을 사용했을 때보다 정확도가 8.89% 향상되었다. 제안된 인과관계 추출 모델은 문장 간의 인과관계 추출에서도 74.67%의 정확도를 보였다.

결 론

본 연구에서는 문헌에서 인과관계를 추출하기 위한 학습 방법을 제안하였다. 제안한 학습 방법은 인과관계 단서구문과 어휘 쌍 확률을 같이 사용한다. 제안된 비지도식 학습 방법을 통해 학습된 인과관계 분류기는 기존의 학습 방법에 비해 6.32% 높은 정확도를 보였다. 이러한 결과는 지도식 학습 방법을 통한 인과관계 분류기[3]와 유사한 성능이다.

제안된 분류 모델은 인과관계 부착 코퍼스를 사용하지 않는 비지도식 학습 방법을 사용하므로, 새로운 영역에 대한 모델의 이식이 용이하다. 또한, 사전에 근거하는 분류 방법과는 달리 사전 미등록어에 의한 오류발생이 없으므로, 질병 분야와 같은 전문분야 문헌에 대한 인과관계 추출 시스템으로도 사용될 수 있다. 마지막으로, 제안된 분류 모델은 명사구 간, 문장 간에 존재하는 인과관계의 추출에 두루 적용할 수 있으며, 단서구문에 따라 명사구와 문장 간, 명사구 내 명사 간의 인과관계 추출에도 사용될 수 있다.

어휘의 의미 개념 쌍 확률을 같이 사용하여 제안된 분류 모델의 확장도 가능하나, 어휘의 의미 애매성을 해결하는 것이 선결 과제이다. 또한 제안된 분류 모델의 완전 자동 학습을 위해서는 단서구문의 자동 확장과 신뢰도의 자동 학습 과정 등이 풀어야 할 문제이다. 또한 이와 더불어 추출된 인과관계의 질의응답 적용은 실험을 통해 증명할 문제이다.

구현된 인과관계 분류기는 인과관계 질의응답에 사용될 수 있으며, 주어진 단어에 대한 인과관계 브라우징 및 요약에도 사용이 가능하다.[15]

REFERENCES

1) Moldovan DS, Harabagiu R, Girju P, Morarescu F, Lacatusu A, Novischi A, Badulescu O(2002) : *Bolohan,"LCC Tools for Question Answering," in Proceedings of the TREC-11 conference, NIST*
2) Girju R, Moldovan D(2002) : *"Mining Answers for Causation Questions," in AAAI Symposium on Mining Answers from Texts and Knowledge Bases*
3) Girju R(2003) : *"Automatic Detection of Causal Relation for Question Answering," Workshop in ACL-03*
4) Modovan D, Pasca M, Harabagiu S, Surdeanu M(2003) : *"Performance Issues and Error Analysis in an Open-Domain Question Answering," ACM Transactions on Information Systems, Vol. 21, No. 2, April, pp133-154*
5) Burger JC, Cardie V, Chaudhri R, Gaizauskas S, Harabagiu D, Israel C, Jacquemin C-Y, Lin S, Maiorano G, Miller D, Moldovan B, Ogden J, Prager E, Riloff A, Singhal R, Shrihari T, Strzalkowski E, Voorhees R. Weishedel, "Issues, tasks, and program structures to roadmap research in question & answering," 2001, http://www-nlpir.nist.gov/projects/duc/roadmapping.html
6) Voorhees E(1999) : *"The TREC-8 Question Answering track report," Proceedings of the TREC-8 conference, NIST*
7) Marcu D, Echihabi A(2002) : *"An Unsupervised Approach to Recognizing Discourse Relations," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002) Conference. Philadelphia, PA, July, pp7-12*
8) Sidner C(1986) : *"Focusing in the comprehension of definite anaphora," Readings in Natural Langugae Processing, Morgan Kaufmann ED, pp363-394*
9) 윤평현(1989) : 국어의 접속어미 연구. 한신문화사

15 http://gensum.kaist.ac.kr/~dschang/ENC에 주어진 문장에서 인과관계를 추출하는 시연 시스템이 구축되어 있으며, 또한 주어진 주제에 대한 인과관계 브라우징 시연이 가능하다.