한국어 특허문서상에서의 인과관계 관찰 및 추출

이신목 김현수 황금하 최기선
KAIST 전자전산학과
{smlee, hyunshu.kim, hgh}@world.kaist.ac.kr, kschoi@cs.kaist.ac.kr

Pattern-based Extraction of Causal Relations from Korean Patent Documents with Two Types of Criteria

Sheen-Mok Lee Hyun-Shu Kim Jin-Xia Huang
Department of EECS, Korea Advanced Institute of Science and Technology

요 약

인과관계는 인간의 인지활동에 있어서 매우 중요한 역할을 한다. 특히 과학과 공학 분야에서 얻은 인과지식은 해당 분야를 이해하는 데에 중요한 역할을 한다. 대표적인 예로, 이들 분야 문서들의 논리적 흐름을 파악하는 데 사용 가능하다. 본 연구에서는, 정보기술 분야의 특허 문서들로부터 얻은 인과 지식을 획득하기 위하여, 문장 내에 나타나는 인과쌍들을 추출하는 방법론을 제시한다. 이를 위하여, 인과관계를 수동으로 태깅하고 관찰하는 작업을 수행하였으며, 태깅을 위한 기준을 설정하였다. 인과쌍의 추출은 패턴을 이용하여 수행하였다.

1. 서론

인과관계(causality)는 그 자체로서 정확하게 정의하기 매우 힘든 개념이지만, 일반적으로 영어 사전에서는 causality를 "the relation between a cause and effect or between regularly correlated events"로 정의한다.

인과관계는 인간이 목표를 향한 노력을 지속하는 데 충분한 두뇌활동이다[5]. 특히, 인과지식은 과학 분야에 있어서 필수적인 요소이다[6]. 예를 들어, 인과지식은 과학과 공학 분야에 있어서 문서의 논리적 흐름을 이해하고 평가하는 데에 사용 가능하다.

본 연구의 최종 목표는 정보기술 분야의 인과지식을 사용하여 정보기술 분야의 특허 문서 상의 인과적 흐름을 추출하고, 논리적 일관성을 평가하는 것이다. 정보기술 분야의 인과 지식을 생성하기 위하여, 특허 문서로부터 인과쌍을 추출하는 방법을 확립할 필요가 있다. 본 논문에서는, 이와 같은 인과쌍을 추출하는 방법론으로서, 기 추출 인과 패턴을 이용한 방법을 제시한다.

문서로부터 추출 가능한 인과관계는 크게 두 가지의 타입으로 분류 가능하다. 인과적으로 연관된 사건들이 문서상에서 연속적으로 나타나는 경우를 미시적 인과관계라 한다. 한편, 인과 추출 과정을 통하여 얻은 서로 인접하지 않은 인과관계를 거시적 인과관계라 한다. 본 논문에서는, 미시적 인과관계에 초점을 맞추고자 한다.

본 논문에서는, 한국어 특허 문서로부터 인과관계를 관찰하고 추출하는 과정을 보이고자 한다. 한국어에 나타나는 사건들은 명사구의 형태로 나타나는 영어에서와는 달리, 주로 동사구나 절의 형태로 나타난다는 특징이 있다. 따라서, 본 논문에서는 명사구간 인과관계 뿐 아니라, 모든 가능한 사건들 간의 인과관계를 추출 대상으로 삼는다.

한국어에서의 인과사건들은 주로 인과관계를 나타내는기능어를 사이에 두고 연결되어 있는 경우가 많다. 본 논문에서는 이들을 인과기능어라고 정의하며, 학습을 통하여 추출한 인과기능어 패턴을 이용하여 인과관계를 추출하였다.

인과기능어 패턴을 추출하기 위하여서는 인과관계 부착 말뭉치가 필요하다. 하지만, 인과관계를

판별하는 것은 쉽지 않은 작업이므로, 인과관계 부착작업을 수행하려면, 인과관계 판별을 위한 기준이 반드시 필요하다.

본 연구에서는, 인과기준을 설정하고, 특허 문서 말뭉치에 인과관계를 부착하였으며, 이로부터 학습한 인과기능어 패턴을 이용하여 인과관계를 추출하였다. 추출 대상은 한국어의 특성을 반영하여, 모든 형태의 사건간의 쌍(명사구, 동사구, 절)을 대상으로 하였다.

본 논문은 다음과 같이 이루어진다. 2장에서 영어와 한국어를 대상으로 한 인과관계의 추출에 관한 기존 연구들을 살펴보고, 3장에서, 인과관계 기준설정 및 수동 말뭉치 부착 작업을 설명한다. 4장에서, 자동 추출 방법론을 소개하고, 5장에서 실험 결과를 보인 후, 6장에서 결론을 맺는다.

2. 관련연구

인과관계 추출에 관한 연구는 현재까지 주로 영어권에서 수행하여 왔다. [4]는 패턴 기반 방법을 이용하여 인과관계를 추출하였다. 그러나, 단순한 패턴 기반 알고리즘으로는 만족할 만한 결과를 얻기 힘들었다.

[3]은 사건명사구에 나타나는 단어들의 시소러스 정보를 사용하였다. [2] 역시 유사한 방법을 사용하였지만, 어휘 및 의미 제약조건을 추출하기 위하여 C4.5 학습 알고리즘을 사용하였다. 패턴 매칭 기법과 함께 어휘 및 의미 자질을 사용함으로써, 성능 향상을 기할 수 있었다. 그러나, 그들의 방법은 시소러스 상에 포함되지 않은 단어들로 이루어진 사건의 추출에는 도움을 주지 못한다는 약점이 있다.

[1]은 인과관계 추출 작업을 분류의 문제로 상정하여, 인과관계 여부를 판단하기 위하여 이진 분류기를 사용하였다. Naïve Bayes 분류기가 사용되었으며, 단서구문 패턴, 어휘쌍 확률, 개념쌍 확률들을 자질로 사용하였으며, 영어 인과관계 추출에 있어서 가장 높은 성능을 보였다. [1]은 또한 한국어 문서로부터 문장간 인과관계를 추출하였다. 그들이 명사구 간 인과관계가 아닌 문장간 인과관계를 추출한 이유는 한국어에서는 사건이 명사구보다는 동사구나 절의 형태로 나타나는 경우가 많기 때문이다[1]. 그들은 추출의 대상이 되는 관계를 접속형태에 따라서, 인과관계, 결과관계, 조건관계로 구분하였다.

그러나, 인과관계 추출을 위한 과거의 연구들에서는 인과관계를 어떠한 방식으로 정의할 것인가에 대한 구체적 기준을 제시하지는 않았다. 하지만, 본 연구에서는 특정 문서 집합 상에서의 인과 추론을 최종 목표로 삼고 있으므로, 인과관계를 정확하게 정의할 필요가 있다.

본 논문에서는, 인과관계 수동부착 과정에서 얻어진 인과관계 판단기준의 대략적인 구조를 제시하고 이를 바탕으로 기본적인 인과관계 자동 추출 과정을 보이고자 한다.

3. 인과관계 수동부착

미시적 인과관계의 자동 추출을 위하여서는 자질의 학습에 사용할 인과 부착 말뭉치가 필요하다. 본 연구에서는, 300개의 정보통신 분야 특허 문서로부터 458개의 인과관계를 수동으로 부착하는 작업을 수행하였다. 한 사람이 수동 부착 작업의 초안을 제시하면, 이를 세 사람이 평가하는 방식으로 진행하였으며, 모든 참여자는 정보통신분야의 전문가로 이루어졌다. 한편, 수동부착 작업과 더불어 인과관계 기준의 설정 작업이 이루어졌다.

3.1. 인과관계 태깅기준의 설정

현재까지의 연구에서, 두 개의 사건 사이에 인과관계가 존재하는지의 여부를 결정하기 위한 기준은 설정된 바가 없다. 인과 기준의 설정 문제를 해결하기 위하여, 인과관계 부착 과정과 더불어 인과기준의 설정 작업을 진행하였다.

인과관계의 초기 기준은 초기 부착자에 의하여 수행되었다. 초기 기준은 세 사람의 평가자가 확인 작업을 하였고, 초기 부착자 및 평가자들 간의 토론을 통한 수정작업을 거쳤다.

인과기준은 예제를 중심으로 나열하였다. 즉, 인과관계 판단 여부가 애매한 모든 예문을 나열하고, 이들에 대하여 어떠한 방식으로 판단하였는지에 대한 기준을 제시하였다.

본 논문에서는 인과관계의 존재 여부 판단을 위한 기준과 존재한다고 가정하였을 경우의 인과관계 범위 설정에 대한 기준의 두 가지 유형으로 인과기준을 구분하였다.

3.1.1. 인과관계 존재여부 판단을 위한 기준

본 절에서는 인과관계 존재여부의 판단과정에서 애매성이 존재하는 대표적 예제들을 나열하고,

각각의 경우에 대한 판단기준을 설정한다.

(1) 추가 전제 필요 여부
두 개의 사건이 인과관계로 연결되기 위하여 추가 전제가 필요한 경우, 인과관계로 취급하지 않는다. 문장 (1)은 인과관계의 판단을 위하여 영역 지식에 기반한 추가 전제가 필요한 경우이다.

[기입 동작이 벡터를 오버라이트]+하기 때문에, [벡터를 오버라이트 전에 독출]+해야 한다. (1)

문장(1)에서의 두 사건,"기입 동작이 벡터를 오버라이트하기 때문에,"와 "벡터를 오버라이트 전에 독출해야 한다." 사이의 관계를 인과관계로 볼 수 있는지 결정하여야 한다. 두 사건은 인과관계를 나타내는 기능어열 "하기 때문에,"로 연결되어 있어서, 인과관계라고 생각할 수 있다. 하지만, 두 사건 사이에는 다른 전제, 즉, "벡터를 오버라이트하면 데이터를 상실한다."는 사실이 뒷받침되어야 올바른 이해가 가능하다. 따라서, 문장(1)에 나타난 두 사건 사이의 인과관계는 인정하지 않는다.
문장(1)과는 달리, 문장(2)는 추가적인 전제를 필요로 하지 않는 명백한 인과관계를 보인다.

[수은]+으로 인해 [전기적으로 활성화]+된다. (2)

(2) 의미적 관계를 기반으로 판단
인과관계를 판단하는 데 있어서 중요한 것은 두 사건 간이 인과적인 표현으로 연결되어 있다는 것이 아니라, 두 사건 간의 의미적인 인과관련성이다. 문장(3)은 도구관계로 연결되어 있지만, 의미적으로는 인과적으로 연결된 사건쌍의 예이다.

[음절톤을 압축]+함으로써 [정보량을 감소]+시킨다. (3)

문장(3)에서는 "음절톤을 압축"하는 사건과 "정보량을 감소"하는 사건 간에 의미적 인과관계가 성립한다. 이와 같이, 사건 간의 연결관계에 상관없이, 의미관계를 기준으로 하여 인과관계를 판단한다.
이외에도 다양한 연결관계에 대하여 인과관계의 존재 여부를 기준으로 설정하였다.

3.1.2. 인과관계 범위설정을 위한 기준
3.1.1의 기준을 이용하여 주어진 문장에 인과관계가 존재한다고 결정한 경우, 다음으로 해결하여야 할 문제는 원인/결과 사건 및 연결부의 범위를 지정하는 일이다. 본 절에서는 사건의 범위 지정과 연결부의 범위 지정으로 나누어서 설명한다.

(1) 사건의 범위 지정
원인 및 결과 사건의 범위를 지정하는 일은 각 사건의 중심어휘를 결정하는 방법과 중심어휘의 좌측에 위치한 수식어구를 결정하는 밥법으로 나뉜다.
중심어휘 선정에 있어서의 애매성이 존재하는 경우, 애매성이 존재하는 후보들 가운데 가장 왼쪽에 위치한 의미 단위를 선정한다.

[?이와 같은 [?렌즈의 [?초점거리단축]+에 의해, [광처리장치의 소형화?]+를 도모?]+한다. (5)

문장(5)의 결과 사건은 "광처리장치의 소형화를"로 표현하는 방법과 "광처리장치의 소형화를 도모한다."로 표현하는 방법이 존재한다. 기술한 중심어휘 선정의 원칙에 의하여, "광처리장치의 소형화를"을 결과사건으로 선택한다. 이와 같이, 최좌측 의미 단위를 중심 어휘로 선정하는 이유는, 동일한 의미를 전달하는 사건 표현 가운데 가장 간결하기 때문이다.
한편, 수식어구의 선정에 있어서 애매성이 존재하는 경우, 두 가지 조건에 의하여 선정한다. 첫째, 특정 수식어구 부분의 의미를 이해하기 위하여, 주어진 문장 이외의 문맥 정보가 필요한 경우, 해당 부분을 수식어구에서 제외한다. 문장(5)에서 "이와 같은"이 그와 같은 예이다. 둘째, 첫번째 조건에서 제외되지 않은 후보들 가운데 가장 긴 수식어구를 선정한다. 문장(5)의 경우, "렌즈의 초점거리단축에 의해"와 "초점거리단축에 의해" 가운데 "렌즈의 초점거리단축에 의해"를 선정한다.

(2) 연결부의 범위 지정

한국어 인과관계는 일반적으로 다음과 같은 두 가지 접속형태를 지닌다.

<원인사건> <연결부> <결과사건> (a)
<원인사건> <연결부1> <결과사건> <연결부2> (b)

예를 들어서, 문장(2)에서는 "으로 인해"가 연결부의 기능을 하므로 (a) type의 접속형태를 지니며, 문장(5)는 "에 의해,"와 "도모한다."가 연결부의 역할을 하므로 (b) type의 접속형태를 지닌다.

3.2. 수동 부착 작업과정

그림1은 인과관계의 수동 부착 과정을 보인다.

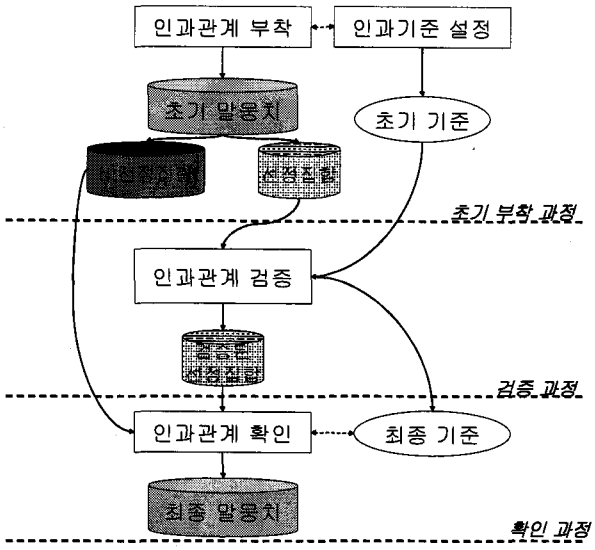

그림1. 인과관계의 수동 부착과정

우선, 정보통신 분야의 전문가인 인과관계 초기 부착자가 구문분석 말뭉치 상에 나타난 인과관계를 수동으로 추출한다. 초기 부착 과정 이전에는 인과관계의 인과관계 태깅 기준을 올바르게 설정할 수 없으므로, 인과관계 부착과 태깅 기준의 설정 과정은 동일인에 의하여 동시에 이루어진다.

초기 부착자가 인식한 인과관계가 부정확할 수 있으므로, 세 사람의 다른 정보통신분야 전문가를 평가자로 선정하였다. 평가 과정은 검증과 확인 과정으로 구분한다. 검증 과정에서는 초기 부착 결과의 일부를 선정집합으로 샘플링하여 초기 부착 기준과 함께 검토 대상으로 삼는다. 즉, 이 과정에서는 인과관계 부착 뿐 아니라, 부착기준 자체도 평가 대상이 된다. 확인 과정에서는 전체 인과관계 부착 결과를 대상으로 검증집합을 통하여 확정된 최종 기준에 부합하는지의 여부를 판단한다. 각각의 단계에서, 평가자들이 코멘트한 결과는 초기 부착자와의 토론 과정을 거쳐서 최종 판단을 내렸다.

3.3. 수동 부착 결과

초기 부착된 인과관계의 수는 468개이다. 초기 부착자는 태깅과정에서 주관적인 태깅 기준을 적용한다.

검증과 확인 단계에서, 평가자들은 부착된 인과관계 각각에 대하여 점수를 부여한다. 점수 부여 기준은 표1과 같다.

표1. 수동부착 결과에 대한 평가점수 부여기준

점수	기준
3	인과관계가 주어진 문장에 존재하며, 인과구성요소의 범위가 정확
2	인과관계가 주어진 문장에 존재하며, 인과구성요소 중 하나의 범위가 부정확
1	인과관계가 주어진 문장에 존재하며, 인과구성요소 중 두 개의 범위가 부정확
0	인과관계가 존재하며 모든 구성요소의 범위가 부정확하거나, 인과관계가 존재하지 않는다.

평가자가 주어진 문장 내의 인과관계 존재 여부에 대하여 확신하지 못하는 경우, 평가자와의 토론을 거쳐서 1점이나 2점을 부여하였다. 세 사람의 점수 부여 결과를 합쳐서 9점 만점의 총점을 산출하였다.

검증 과정에서는 162개의 인과관계를 선정집합으로 선택하였다. 평가자들은 초기 태깅기준을 자신의 기준에 맞게 수정하였다. 초기 부착자와의 토론 과정을 거쳐서, 8개의 삼진관계가 인과관계 집합에서 탈락하였고, 2개가 새로 부착되었다.

확인 단계에서는, 전체 말뭉치를 대상으로, 확정된 태깅 기준에 맞는지를 확인하였다. 이 단계 이후, 4개의 삼진관계가 제거되었고, 10개가 내부구성요소 범위변경을 하였다.

이와 같은 과정을 거쳐서, 최종적으로 458개의 인과관계를 얻었다. 확인 과정에서 전체 인과관

계 후보 집합에 대하여 부여한 총점의 분포는 그림2와 같다.

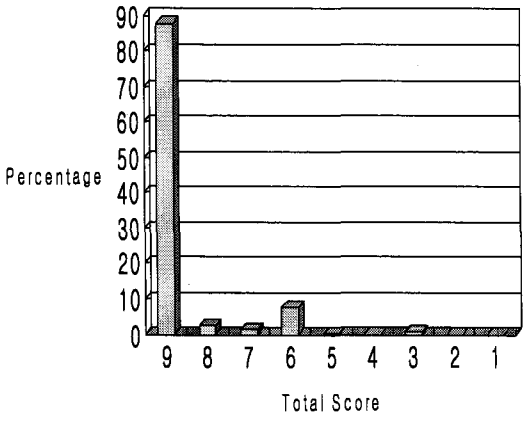

그림2. 수동부착 결과에 대하여 확인단계에서 부여한 총점 분포

위의 분포에서 알 수 있듯이, 모든 평가자가 수동 부착자의 의견에 동의하는 비율(9점)은 88%였다. 즉, 해당 영역의 인과관계 자동추출 작업 성능의 상한선은 88%이다.

4. 패턴 기반 인과관계 추출

본 장에서는 3장에서 얻은 인과관계 부착 말뭉치를 이용하여 얻은 패턴을 기반으로 한 자동 추출 작업을 소개한다. 본 연구에서 인과관계 추출을 위하여 주로 사용한 패턴은 인과기능어이다. 인과기능어는 인과관계를 표현하기 위한 기능어를 뜻하며, 일반적으로 연결부의 기능어부에 의하여 표현된다.

예를 들어서, 문장(1)의 "하기 때문에"는 절 간의 인과관계를 표현하는 기능을 하므로, 인과기능어이다. 인과기능어는 일반적으로, 원인과 결과 사건 사이에 위치하므로, 인과관계 추출에 유용한 자질이다.

인과기능어 자질의 추출과정은 그림3과 같다. 우선, 인과부착 말뭉치로부터 얻은 각 인과기능어 후보에 대하여, 단어별 신뢰도를 식(1)의 방법으로 계산한다.

$$causal\_word\_relib(f_i) = \frac{causal\_word\_freq(f_i)}{word\_freq(f_i)} \quad (1)$$

식(1)에서 $word_freq(f_i)$는 인과기능어 후보 f_i의 빈도수이고, $causal_word_freq(f_i)$는 f_i가 인과관계 연결부의 역할을 하는 빈도수이다. 이때의 모든 빈도수 정보는 인과부착 말뭉치로부터 추출한다.

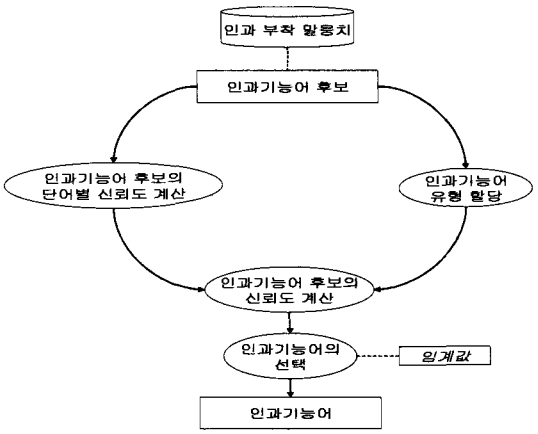

그림3. 인과기능어의 학습과정

인과기능어 후보의 단어별 신뢰도는 인과기능어 판별에 중요한 역할을 하지만, 말뭉치에 부착된 인과관계의 개수가 많지 않아서, 데이터 희귀성의 문제가 나타난다는 문제점이 있다. 따라서, 본 연구에서는 인과기능어의 유형이라는 개념을 도입하여, 유형별 신뢰도를 계산하였다. 인과기능어 유형은 각 인과기능어 후보에 대하여, 인과관계를 가장 잘 표현하는 형태소로 정의한다. 예를 들어, "하기 때문에"라는 인과기능어 후보는 "하","기","때문","에"의 네 개의 형태소로 나눌 수 있는데, 이 가운데 인과관계를 표현하는 데 가장 중요한 형태소는 "때문"이다. 따라서, "하기 때문에"의 인과기능어 유형은 "때문"이 된다. 인과관계 유형의 할당을 위하여, 다음과 같은 식을 사용한다.

$$causal\_type\_relib(m_j) = \frac{causal\_type\_freq(m_j)}{type\_freq(m_j)} \quad (2)$$

$$type(f_i) = \arg\max_{1 \le j \le n} causal\_type\_relib(m_j) \quad (3)$$

식(2)에서는 인과기능어 후보에 포함된 형태소 m_j에 대한 유형 신뢰도를 계산한다. $type_freq(m_i)$는 인과기능어 후보 m_i의 빈도수이고, $causal_type_freq(m_i)$는 m_i가 인과관계 연결부에 포함되는 빈도수이다. 식(3)에서는 식(2)에서 얻은 각 형태소의 유형 신뢰도를 이용하여, 인과기능어 $f_i(=m_1m_2...m_n)$의 유형을 계산한다. 최종적인 인과기능어 신뢰도는 식(4)의 방식으로 계산한다.

$$causal\_relib(f_i) = causal\_word\_relib(f_i) \bullet causal\_type\_relib(type(f_i)) = \frac{causal\_word\_freq(f_i) \bullet causal\_type\_freq(type(f_i))}{word\_freq(f_i) \bullet type\_freq(type(f_i))} \quad (4)$$

인과기능어 신뢰도가 수동으로 지정한 임계값을 넘어서는 경우, 인과기능어로 선택하였다. 이와 같이 학습한 인과기능어 패턴을 말뭉치에서 탐색하여, 바로 앞과 뒤에 나타나는 사건을 추출하였다. 사건 추출은 구문 분석 결과를 사용하였으며, 휴리스틱을 이용하여 사건의 범위를 한정하였다.

5. 실험 및 분석

본 연구에서는, 웹사이트로부터 얻은 300개의 문서로 이루어진 정보통신 분야의 특허 말뭉치를 사용하였다. 이 말뭉치로부터 인과기능어의 자질을 학습하였으며, 같은 말뭉치에서 평가를 수행하였다. 추출된 인과기능어의 수는 40개이다. 표2는 이들 가운데 높은 점수가 매겨진 일부를 보인다.

표2. 높은 순위의 신뢰도를 기록한 인과기능어

인과기능어	인과유형	인과신뢰도
하기 때문에	때문	0.1649
되지 않기 때문에	때문	0.1484
어지기 때문에	때문	0.1237
수 있기 때문에	때문	0.1237
되기 때문에	때문	0.1164
기 때문에	때문	0.1016
이므로	므로	0.0999
하기 때문에	때문	0.0989

한국어 문서에서 인과관계로 나타나는 사건들은 명사구보다는 절의 형태로 나타나는 경우가 많아서, 비교적 길이가 길고, 범위의 지정이 어렵다. 따라서, 인과기능어 패턴을 잘 추출하였다고 해도, 사건 범위를 잘못 지정하여 인과관계 추출이 실패하는 경우가 많다. 따라서, 성능 평가에 있어서도, 인과관계 자체의 유무를 판단하는 작업에 대한 평가 이외에, 인과관계가 제대로 추출된 경우에 한하여 사건의 범위를 지정하는 작업에 대한 평가를 따로 수행하였다.

인과관계 자체의 유무를 판단하는 작업에 대한 평가를 위하여, 인과기능어 탐색 성공 여부에 대한 평가를 수행하였으며, 사건 범위의 지정을 평가하기 위하여, 시작 부분의 추출과 끝 부분의 추출 성공 여부를 평가하였다. 이를 위하여, 표3과 같은 정답인과쌍과의 유사도 점수를 추출된 각각의 인과쌍에 부가하였다.

표3. 추출 인과쌍과 정답인과쌍의 유사도 점수

시작 표식 추출	끝 표식 추출	부가 점수
정확한 추출	정확하게 추출	1
정확한 추출	부정확한 추출	0.5
부정확한 추출	정확한 추출	0.5
부정확한 추출	부정확한 추출	0

표3에서 추출한 유사도 점수를 식(5)에 적용하여 사건 범위의 정확률을 계산한다.

$$P = \frac{sum\ of\ similarity\ scores\ of\ extracted\ relations}{\#\ of\ correctly\ extracted\ causal\ relations}$$

(5)

이와 같은 방법을 이용하여 성능을 계산한 결과는 표4와 같다. 본 시스템의 성능이 아직 매우 낮은 가장 큰 원인은 패턴 매칭 기법의 한계 때문이다. 본 시스템에서 추출한 인과기능어 패턴 이외에도 많은 패턴들이 수동 말뭉치에 존재하며, 이들을 올바르게 추출하려면 사건 간의 의미 관계를 반드시 고려하여야

표4. 인과관계의 추출과 사건범위 추정의 성능평가결과

인과관계 추	정확율	56.46%
	재현률	17.73%
사건 범위 추정 정확율		90.36%

한다. 두번째 원인은 표2에서 알 수 있듯이 인과기능어 패턴의 신뢰도가 매우 낮다는 점이다. 인과기능어 패턴의 신뢰도가 현격히 낮은 이유는 인과관계 부착 과정에서 인과기준을 엄격하게 적용하였다는 점을 들 수 있다. 특허 분야의 인과관계가 영역 지식이 충분하지 않은 상태에서 쉽게 이해하기 힘들며, 이해할 수 있는 인과관계만을 우선적인 태깅 대상으로 삼았기 때문이다.

6. 결론

본 연구에서는, 한국어 특허 문서에 나타난 인과관계를 관찰하여 인과 기준을 설정하였다. 이와 같은 기준에 따라, 인과관계 부착 말뭉치를 생성하였으며, 인과관계를 표현하는 기능어의 패턴을 이용한 자동 추출을 수행하였다. 한 문장 내에 나타나는 모든 구문 단위의 인과쌍을 추출 대상으로 하였다.

아직 충분한 수행도를 보이지 못하고 있는 이유는 기능어가 추출자질로서 충분하지 못하다는 점과 수동으로 구축한 말뭉치 량이 충분하지 못하다는 점을 들 수 있다.

향후 연구로서, 사건 내의 내용어로 자질 집합을 확장할 예정이며, 말뭉치와 자질의 량을 늘리기 위하여 EM 알고리즘을 적용하고자 한다. 또한, 본 연구에서 대상으로 한 "확실한" 인과관계 이외에 인과추출을 위하여 필요한 관계의 유형을 정리하고자 한다.

참고 문헌

[1] 장두성. 분류 기법과 단서구문 학습모델을 이용한 인과관계 지식추출 자동화 연구. 한국과학기술원 박사학위논문, 2005.

[2] Girju, Roxana. Automatic Detection of Causal Relation for Question Answering. Workshop in the 41st Annual Meeting of the Association for Computational Linguistics Conference(ACL-03). 2003.

[3] Girju, Roxana and Moldovan, Dan. Mining Answers for Causation Questions. AAAI Symposium on Mining Answers from Texts and Knowledge Bases. 2002.

[4] Khoo, Christopher S. G. et al. Automatic Extraction of Cause-Effect Information from Newspaper Text without Knowledge-based Inferencing. Literary and Linguistic Computing. Volume 13, Number 4, pp. 177-186. 1998.

[5] Minsky, Marvin. The Society of Mind. New York:Simon and Schuster. 1986.

[6] Mackie, John. Causes and Conditions. American Philosophical Quarterly, 2(4):245-264.

[7] Russell, Bertrand. Human Knowledge. New York:Simon and Schuster. p. 487. 1948.

[8] Merriam-Webster Online Dictionary. http://www.m-w.com. 2005.