韓國컴퓨터情報學會 論文誌
第17卷 第2號, 2012. 2.

2012-17-2-4-1

텍스트 마이닝 및 자동 추론 기반 생물학 지식 발견 시스템을 위한 확률 기반 필터링

이 희 진*, 박 종 철*

Probabilistic filtering for a biological knowledge discovery system with text mining and automatic inference

Hee-Jin Lee*, Jong C. Park*

요 약

본 논문에서는 텍스트 마이닝을 통해 생물학 문헌에서 분자 수준의 사건(event) 정보를 자동으로 추출하고, 이들 사건 정보를 기반으로 새로운 생물학 지식을 자동 추론하는 텍스트 마이닝 - 추론 통합 구조의 시스템을 다룬다. 이러한 통합 구조의 지식 발견 시스템은 미리 추출되어 데이터베이스에 등록된 정보만을 입력으로 사용하는 시스템들에 비하여 최신 정보를 보다 빨리 사용할 수 있고, 미리 정의된 형식 이외의 다양한 정보를 사용할 수 있다는 장점이 있다. 반면, 텍스트 마이닝 정보 추출 결과를 그대로 사용하기 때문에 텍스트 마이닝 모듈(module)의 성능에 따라 전체 시스템의 효용성이 크게 저하될 수도 있다는 문제가 있다. 본 논문에서는 확률 기반 필터링(filtering) 방법을 제안하여, 텍스트 마이닝 결과 중 양성 오류(false positive)를 효과적으로 제거함으로써 전체 지식 발견 시스템의 정확도 및 효용성을 높이고자 한다. 본 논문에서 제안한 확률 기반 필터링 방법은 기준(baseline) 방법으로 사용된 횟수 기반 필터링 방법보다 높은 성능을 보였다.

▸ Keyword : 지식 추론 시스템, 텍스트 마이닝, 자동 추론, 확률 기반 필터링

Abstract

• 제1저자 : 이희진 • 교신저자 : 박종철
• 투고일 : 2011. 11. 08, 심사일 : 2011. 11. 17, 게재확정일 : 2011. 11. 22.
* 카이스트 전산학과(Dept. of Computer Science, KAIST)
※ 이 논문은 정부(교육과학기술부)의 재원으로 한국연구재단의 지원(No. 20110029447) 및 한국학술진흥재단의 지원을 (KRF-2007-313-D00738) 받아 수행된 연구임.

In this paper, we discuss the structure of biological knowledge discovery system based on text mining and automatic inference. Given a set of biology documents, the system produces a new hypothesis in an integrated manner. The text mining module of the system first extracts the 'event' information of predefined types from the documents. The inference module then produces a new hypothesis based on the extracted results. Such an integrated system can use information more up-to-date and diverse than other automatic knowledge discovery systems use. However, for the success of such an integrated system, the precision of the text mining module becomes crucial, as any hypothesis based on a single piece of false positive information would highly likely be erroneous. In this paper, we propose a probabilistic filtering method that filters out false positives from the extraction results. Our proposed method shows higher performance over an occurrence-based baseline method.

▸ Keyword : knowledge discovery system, text mining, automatic inference, probabilistic filtering

I. 서 론

생물학 분야에서는 방대한 지식의 양과 그 빠른 증가 속도로 인해 자동화된 기법의 도움을 받아 문헌, 데이터베이스 등에 분산되어 보관되어 있는 관련 정보들을 효율적으로 수집하여 지식 발견에 활용하고자 하는 요구가 있다. 이에 따라 문헌에 나타난 특정 정보들을 텍스트 마이닝(text mining) 기법을 이용하여 추출하는 연구가 많이 진행되고 있다[1]. 최근에는 관련 정보를 자동으로 추출하여 수집하는 단계를 넘어서 수집된 정보를 통합하고 관리하는 자동화된 시스템의 필요성이 대두되고 있는데[2,3], 이는 관련 정보를 자동화된 방법으로 수집하더라도 개별 연구자가 방대한 양의 정보를 종합적으로 파악하여 이를 바탕으로 새로운 지식을 발견하는 데에는 아직 어려움을 겪고 있기 때문이다.

이에 따라 본 논문에서는 텍스트 마이닝을 통하여 생물학 문헌에서 분자 수준의 사건(event) 정보를 자동으로 추출하고, 이들 사건 정보를 기반으로 새로운 생물학 지식을 자동 추론하는 시스템의 구성에 대해 논의한다. 본 논문에서 제시하는 생물학 지식 자동 추론 시스템은 텍스트 마이닝과 자동 추론이 통합된 형태이다. 이러한 텍스트 마이닝 - 추론 통합 구조의 지식 발견 시스템은 미리 추출되어 데이터베이스에 등록된 정보를 입력으로 사용하는 방식의 생물학 지식 추론 시스템들에[4] 비하여 최신 정보를 보다 빨리 사용할 수 있고, 데이터베이스에 미리 정의된 형식 이외에 보다 다양한 정보를 사용할 수 있다는 장점이 있다. 반면에 텍스트 마이닝의 정보 추출 결과를 그대로 사용하기 때문에 텍스트 마이닝 모듈(module)의 성능에 따라 전체 시스템의 효용성이 크게 저하될 수도 있다는 문제를 가지고 있다. 특히 텍스트 마이닝 모듈의 정밀성(precision)이 중요한데, 이는 잘못된 텍스트 마이닝을 통해 도입된 양성 오류(false positive)가 자동 추론을 위한 기초 정보로 사용된 경우, 이를 바탕으로 얻어낸 가설들이 부정 될 확률이 높아지기 때문이다.

본 논문에서는 텍스트 마이닝 결과에 대한 확률 기반 필터링(filtering) 방법을 제안하고, 이를 통해 텍스트 마이닝 결과 중 양성 오류를 효과적으로 제거함으로써 전체 지식 발견 시스템의 정확도 및 효용성을 높이고자 한다. 본 논문에서 제안하는 필터링 방법은 텍스트 마이닝 모듈 고유의 정밀도 값과 동일한 정보가 반복되어 추출되는 횟수에 기반하는 방법으로, 다양한 정보 추출(information extraction) 결과에 적용이 가능하다. 따라서 다양한 정보의 사용을 위해 여러 종류의 텍스트 마이닝 모듈을 사용할 것이 예상되는 텍스트 마이닝 - 추론 통합 구조의 지식 발견 시스템에 특히 적합하다. 이러한 확률 기반 필터링 방법은 개별 텍스트 마이닝 모듈의 정밀성 향상을 위한 방법들과 상호보완적으로 전체 지식 발견 시스템의 정확도 향상을 가져올 것으로 기대된다.

본 논문에서 제안한 확률 기반 필터링 방법은 이전 연구들에서[5] 사용된 바 있는 횟수 기반 필터링 방법보다 높은 성능을 보인다. 또한 저자들이 아는 한 본 논문은 정보 추출 결과의 필터링에 대한 정량적 분석을 시도한 최초의 연구이다.

이후 본 논문은 아래와 같이 구성된다. 먼저 Ⅱ절에서 문헌 정보를 바탕으로 생물학 지식을 자동으로 발견하고자 한 관련 연구 및 텍스트 마이닝 결과 중의 오류를 제거하고자 한 관련 연구를 소개한다. 이후 Ⅲ절에서는 텍스트 마이닝 - 추론 통합 구조의 생물학 지식 발견 시스템을 소개하고, 이러한 시

스템에 적용시키기 위한 확률 기반 필터링 기법에 대해 논의한다. 또한 제안하는 확률 기반 필터링 방법을 기준(baseline) 방법과 비교하여 그 성능을 확인한다. 마지막으로 IV절에서는 본 논문에서의 논의를 결론짓는다.

II. 관련 연구

문헌 정보를 이용하여 생물학 지식을 자동으로 추론하고자 한 대표적 연구로는 Swanson의 연구가[6,7,8] 있다. 이는 생물학적 개념 A와 생물학적 개념 B 사이에 문헌에 자주 함께 나타나는 공기 관계가 있고(co-occurrence), 생물학적 개념 B와 생물학적 개념 C 사이에도 공기 관계가 있을 때, A와 C사이에 어떠한 의미상의 연관관계가 존재할 것이라는 가정을 기본으로 한다. Swanson의 연구 및 이러한 가정을 공유하는 다른 연구들을[9,10] 통해서 실제로 유용한 생물학적 가설이 발견되기도 했지만, 이러한 방법들을 통해 추론해내는 가설은 그 정확도가 많이 떨어진다는 단점이 있다.

반면에 생물학 문헌에서 텍스트 마이닝을 통해 좀 더 구체적인 정보를 추출하고, 형식 논리 및 규칙 기반의 추론 기법을 이용해서 새로운 생물학적 가설을 생성하고자 하는 연구들도 있다. 추론 규칙의 집합인 'Discovery pattern'을 이용하여 기존의 약물을 새로운 질병의 치료제로 사용할 수 있는지 여부를 판단하고자 한 연구[11], AnsProlog로 기술된 추론 규칙들을 통해 새로운 약물 간 상호작용(drug-drug interaction)을 찾아내고자 한 연구[12] 및 온톨로지(ontology)와 서술논리(description Logic)을 이용하여 새로운 문자 수준의 사건 정보를 발견하고자 한 연구가[13] 그러한 연구이다.

그러나 이러한 텍스트 마이닝 기반의 지식 추론 연구들은, 텍스트 마이닝 과정의 오류들로 인해 전체 추론 결과의 정확도가 떨어지기 쉽다는 문제점을 가지고 있다. 이를 보완하기 위해Giles와 Wren은[5] 특정 횟수 이상으로 반복되어 추출된 정보만을 사용하는 방법을 사용하였다. 그들은 2회 이상 반복되어 추출된 정보들만을 사용하여 카페인(caffeine)과 체내 다른 물질들 사이의 관계를 파악하고자 하였다. 그러나 추출 정보를 사용하기 위한 기준 횟수는 임의로 결정된 것이며, 기준 횟수의 변화에 따른 성능 변화는 보고되지 않았다. Tari와 공동연구자들은[12] 텍스트 마이닝 결과를 기존의 지식 데이터베이스와 비교하여 검증하는 방법을 제안하였다. 그들은 새로운 약물 간의 상호작용(drug-drug interaction)을 문헌 정보를 통해 자동으로 밝혀내고자 하였는데, 이를 위해 문헌에서 각 약물의 효소로 작용하는 단백질들의 정보 및 각 단백질 효소의 활동을 돕는 전사인자(transcription factor)들의 정보를 추출하였다. 이 때 추출 결과 중의 양성 오류 제거를 위해 추출된 단백질 관련 정보들은 기존의 단백질 데이터베이스와 비교되었다. 이들의 방법은 약물 간의 상호작용과 같이 한정된 도메인(domain)에서만 사용가능하다는 단점이 있으며, 이러한 검증 절차에 따른 성능 변화는 알려지지 않았다.

III. 생물학 지식 발견 시스템을 위한 텍스트 마이닝 결과의 확률 기반 필터링

본 절에서는 먼저 텍스트 마이닝 - 추론 통합 구조의 지식 발견 시스템인 BioDetective를 소개한다. 이후 확률 기반 필터링 방법을 제시하고, 이를 BioDetective의 텍스트 마이닝 모듈에 적용하기 위한 구체적 방법을 논의한다. 마지막으로 BioDetective의 텍스트 마이닝 모듈에 적용된 필터링 방법을 테스트하여 그 성능을 평가한다.

1. 텍스트 마이닝 및 자동 추론 기반 생물학 지식 발견 시스템

1.1 시스템 구조

본 연구에서 다루는 텍스트 마이닝 - 추론 통합 구조의 지식 발견 시스템인 BioDetective는[14] 생물학 문헌들 및 사용자 질의(query)를 입력으로 받아, 사용자 질의로 받은 인과관계(causal relation)가 문헌 정보를 통하여 주어진 정보들을 사용하여 추론 가능한 것인지를 판별하여 알려주는 시스템이다. 그림 1은 이러한 BioDetective의 구조를 보여주고 있다.

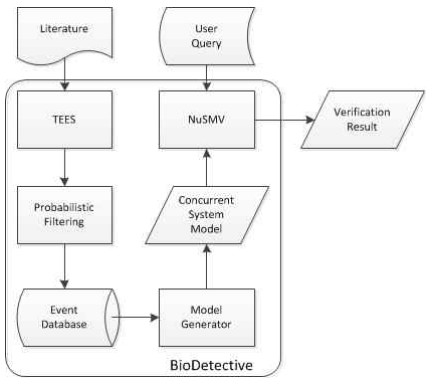


그림 1. BioDetective의 구조
Fig. 1. The structure of BioDetective

BioDetective가 입력으로 받은 생물학 문헌들은 텍스트 마이닝 시스템인 Turku Event Extraction System (TEES)으로[15] 전달되며, TEE는 이들 문헌에 나타난 분자 수준의 사건 정보를 추출하게 된다. 추출된 사건 정보들은 데이터베이스화 되어 모델 생성기(model generator)에 전달된다. 모델 생성기는 데이터베이스의 사건 정보를 활용하여 인과 네트워크를 생성하는데, 이를 병행적 시스템(concurrent system)의 모델로 나타내어 준다. 생성된 병행적 시스템 모델은 다시 NuSMV라는[16] 모델 체커(model checker)의 입력으로 주어진다. 병행적 시스템 모델을 입력으로 받은 NuSMV에 선형 시제 논리 (linear temporal logic) 형태로 기술된 질의문이 주어지면, NuSMV가 병행적 시스템 모델에서 선형 시제 논리 형태로 나타내어진 인과 관계가 성립가능한지 여부를 판단한다. 이 판단 결과가 BioDetective의 최종 출력이 된다.

본 논문에서는 BioDetective의 텍스트 마이닝 모듈인 TEES의 정보 추출 결과를 데이터베이스화하기 전에 확률 기반 필터링 방법을 통해 긍정 오류를 제거하고자 하여, 보다 오류가 적은 정보들을 이후의 병행적 시스템 모델 생성 및 모델 체킹(model checking) 과정에 전달하고, 이를 통해 전체 시스템의 정확도 및 효용성을 향상시키는 방안을 다룬다.

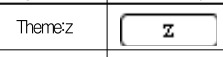
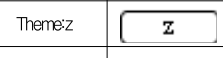
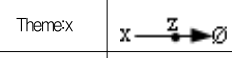
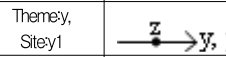
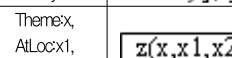
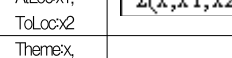
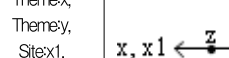
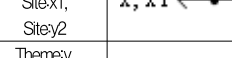
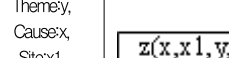
1.2 사건 정보의 구조

BioDetective 내의 텍스트 마이닝 모듈로 사용되는 TEES는 생물학 문헌에서 미리 정의된 형태의 사건 정보를 파악하여 추출한다. 사건 정보는 단백질 상호 작용(protein-protein interaction) 등이 주로 나타내어지는 이진 관계(binary relation)와 달리 좀 더 복잡한 구조를 가지는데, 개별 사건 정보는 사건 유형(type)과 두 개 이상의 인수(argument)로 구성된다. 또한 하나의 사건 정보가 다른 사건 정보의 인수가 될 수 있어 재귀적(recursive) 구조를 보인다. 이러한 특성으로 인하여 사건 정보는 BioDetective의 사건 데이터베이스 내의 정보를 기술하는 'Diagrammatic Pathway Language (DPL)'로[14] 변환될 수 있다.

표 1은 TEES가 추출하는 사건 정보들이 가질 수 있는 사건 유형의 종류 및 인수의 역할(role) 유형들의 종류를 보여준다. 또한 사건 정보의 유형별로 이에 대응되는 DPL 심볼(symbol)들을 함께 보여 주고 있다. 개별 사건 정보가 DPL 심볼로 변환될 때에 사건의 인수들도 함께 DPL 심볼에 표시된다. 표 1에서 사건 유형 및 인수의 역할 유형을 나타내는 문자열에서 콜론(colon) 이후는 해당 사건 정보 및 해당 인수를 나타내는 식별자이다. 이러한 인수 및 사건의 식별자는 DPL 심볼에도 함께 나타나 있는데, 개별 사건 정보가 DPL 심볼로 변환될 때 각 인수들이 심볼 내에서 위치할 부분을 표시하고 있다. 만일 어떤 사건의 인수가 또 다른 사건 정보인 경우, 이에 대응하는 DPL 심볼은 두 개의 DPL 심볼을 연결시켜 놓은 형태가 되며, 이러한 연결이 이어져 하나의 생화학 경로(pathway)를 나타내게 된다.

본 논문에서는 위와 같은 구조로 정의된 사건 정보들 중에서 양성 오류를 효과적으로 필터링 하기 위한 확률 기반 필터링 방법을 제시한다.

표 1. TEES의 사건 유형 및 대응 DPL 심볼
Table 1. Event types used in TEES and corresponding DPL symbols

Event Type	Arguments	DPL Symbol
Gene_expression	Theme:z	
Transcription	Theme:z	
Protein_catabolism:z	Theme:x	
Phosphorylation:z	Theme:y, Site:y1	
Localization:z	Theme:x, AtLoc:x1, ToLoc:x2	
Binding:z	Theme:x, Theme:y, Site:x1, Site:y2	
Regulation:z	Theme:y, Cause:x, Site:x1, CSite:y1	
Positive_regulation	Theme:y, Cause:x, Site:x1, CSite:y1	
Nevative_regulation:z	Theme:y, Cause:x, Site:y1, CSite:x1	

2. 정밀도 향상을 위한 확률 기반 필터링 방법

본 논문에서 제안하는 확률 기반 필터링 방법은 정보 추출 시스템(information extraction system)의 알려진 정밀도 값을 기반으로 개별 추출 결과 및 추출된 정보의 신뢰도를 산정하고, 이를 바탕으로 일정 수준 이상의 신뢰도를 가지는 정보들만을 골라내어 사용하는 방법이다. 본 절에서는 먼저 개별 추출 결과 및 추출된 정보의 신뢰도를 정의하고, 이를 사건 정보에 적용하는 방법에 대해 논의한다. 이어 이러한 신뢰도 정보를 바탕으로 사건 정보를 필터링하는 알고리즘을 제시한다.

2.1 개별 추출 결과 및 추출된 정보의 신뢰도

본 논문에서는 정보 추출 시스템(information extraction system)에 의해 추출된 개별 결과와 이들에 의해 나타내어지는 정보를 분리하여 정의한다. 개별 추출 결과는 개별 문장 또는 문헌에 나타난 정보가 정보 추출 시스템에 의해 확인되어 미리 정의된 구조로 형식화되어 나타내어진 결과이다. 반면에, 추출된 정보는 개별 추출 결과들이 의미하는 정보 그 자체로서, 두 개 이상의 추출 결과가 동일한 정보를 의미할 수 있다. 예를 들어, 어떤 정보 추출 시스템에 의해 다섯개의 서로 다른 문장에서 단백질 'p53'과 'mdm2'가 서로 상호작용을 한다는 정보가 '(p53, mdm2)'라는 형태로 5번 추출되었다고 하자. 이 경우 추출 결과는 다섯개이지만, 추출된 정보는 한 가지이다.

이에 따라 추출 결과의 신뢰도와 추출된 정보의 신뢰도도 분리하여 정의된다. 추출 결과의 신뢰도는 개별 정보 추출 과정에 대한 신뢰도이고, 추출된 정보의 신뢰도는 해당 정보가 지식 발견 시스템의 다음 단계인 자동 추론 시스템의 입력으로 사용될 수 있는지 여부를 판별하기 위한 신뢰도이다. 특히 추출된 정보의 신뢰도는 정보 추출 시스템의 입력으로 사용된 문헌 집합에서 해당 정보가 반복되어 추출된 경우 상승하게 된다. 위에서 예로 들었던 '(p53, mdm2)'라는 결과가 다섯 번 추출 되었을 때보다 10번 추출되었을 때 단백질 'p53'과 단백질 'mdm2'가 상호작용을 한다는 정보 자체의 신뢰도가 상승한다.

즉, 추출 결과의 신뢰도는 정보 추출 시스템에 의해 추출된 임의의 결과가 참긍정(true positive)일 확률로 정의되며, 정보 추출 시스템에 의해 추출된 정보 중 참긍정인 정보의 비율을 나타내는 정밀도 값으로 결정된다. 따라서 정밀도가 p로 알려진 임의의 정보 추출 시스템 A에 의해 추출된 결과 a의 신뢰도는 p가 된다. 반면에 추출된 정보 i의 신뢰도는 문헌에 나타난 모든 정보가 참인 것으로 가정했을 때, 추출된 정보 i가 참일 확률로 정의된다. 만일 정보 i를 나타내는 n개의 추출 결과 $a_i,\ldots,a_n$이 각각의 신뢰도 $p_{a_i},\ldots,p_{a_n}$으로 추출되었다면, 정보 i의 신뢰도 r_i는 아래의 (1)과 같이 결정된다.

$$r_i = 1 - \prod_{j=1}^{n}(1-p_{a_j}) \quad \cdots\cdots (1)$$

위의 (1)에 따라 개별 정보의 신뢰도가 결정되고 나면, 미리 결정된 임계값(threshold) 보다 낮은 신뢰도 값을 가지는 정보를 제외시키는 방법으로 추출된 정보의 필터링을 수행하게 된다.

2.2 사건 정보의 신뢰도

1.2절에서 논의한 것과 같이 사건 정보는 사건의 유형과 인수들로 정의된다. 본 논문에서는 이러한 사건 정보를 추출하는 사건 정보 추출 시스템(event extraction system)의 정밀성 및 재현율(recall)등의 성능(performance) 값을 측정하는데 있어, 두 개 이상의 인수를 가지는 사건 정보를 여러 개의 단일 인수를 가지는 사건 정보로 분할(decompose)하여 분할된 단일 인수 사건 정보들을 기준으로 시스템의 성능을 측정하는 방법을 사용하고자 한다. 즉, 본 논문에서 논의하는 사건 정보 추출 시스템의 정밀도는 시스템에 의해 추출된 전체 사건-인수 쌍(pair) 중에서 올바르게 추출된 사건-인수 쌍의 비율이다. 이러한 성능 측정 방법은 생물학 분야의 사건 정보 추출에 대한 경연인 BioNLP shared task에서 [15] 시도된 방법으로, 보다 복잡한 구조의 사건 정보를 추출하는 것에 가중치를 두는 방법이다. 또한, 이러한 사건-인수의 쌍을 기본 단위로 하는 성능 측정 방법을 사용하면 사건 및 인수의 유형 별로 사건 정보 추출 시스템의 정밀도를 파악하기가 용이한데, 이렇게 세분화된 정밀도 값은 추출 결과 및 추출된 정보의 신뢰도를 보다 세밀하게 측정하는데 사용될 수 있다.

이러한 성능 측정 방법에 기초하여, 사건 정보 추출 시스템 A의 유형 t인 사건의 역할 유형 s인 인수를 추출하는 작업에 대한 정밀도가 $p_{t,s}$일 때, A에 의해 추출된 결과로서 유형 t의 사건 e와 역할 유형 s인 인수 a의 쌍 $e-s$에 대한 신뢰도는 $p_{t,s}$가 된다. 따라서, 단일 인수를 가지는 유형 t의 사건 정보 i를 나타내는 n개의 사건-인수 쌍 $b_i,\ldots,b_n$이 A에 의해 추출되었다면, 정보 i의 신뢰도 r_i는 (1)에 의해 아래의 (2)와 같이 결정된다.

$$r_i = 1 - (1-p_{t,s})^n \quad \cdots\cdots (2)$$

또한, 사건 정보 추출 시스템 A의 유형 t인 사건에 대해 역할 유형 s인 인수를 추출하는 작업에 대한 정밀도가 $p_{t,s}$일 때, A에 의해 추출된 각각 $t_1,\ldots,t_m$의 역할 유형을 가지는 m개의 인수 $a_1,\ldots,a_m$를 가지는 유형 t인 사건 e를 나타내는 단일 추출 결과의 신뢰도는 $\prod_{j=1}^{m} p_{t,t_j}$가 된다. 따라서, m개의 인수를 가지는 유형 t의 사건 정보 i 를 나타내는 n개의 추출 결과 $a_i,\ldots,a_n$이 A에 의해 추출되었다면, 정보 i의 신뢰도 r_i는 (1)에 의해 아래의 (3)과 같이 결정된다.

$$r_i = 1 - (1 - \prod_{j=1}^{m} p_{t,t_j})^n \quad \cdots\cdots (3)$$

2.3 사건 정보를 위한 확률 기반 필터링 알고리즘

본 논문에서 제시하는 사건 정보의 필터링을 위한 확률 기반 알고리즘은 그림 2와 같다. 먼저 개별 사건 정보의 신뢰도 정보를 계산하는데, 이를 위해 해당 사건 정보의 사건 유형 및 인수들의 역할 유형에 따른 사건 정보 추출 시스템의 정밀도 값을 참조한다. 또한 해당 사건 정보를 나타내는 추출 결과가 전체 입력 문헌에서 추출된 횟수를 이용한다. 사건 정보 추출 시스템에 의해 추출된 모든 정보들의 정밀도를 계산하고 나면, 계산된 정밀도가 사용자에 의해 주어진 임계값보다 높은 정보들을 선택한다. 이 때, 다른 사건을 인수로 가지는 재귀적 구조를 가진 사건 정보의 경우, 인수가 되는 사건 모두의 정밀도가 임계값보다 높아야만 선택될 수 있다. 또한, 이러한 기준은 인수가 되는 사건이 또다시 재귀적인 구조를 가지는 경우에도 반복되어 적용된다.

그림 2의 알고리즘에서 사용자에 의해 지정된 임계값은 (3)에 의해 계산된 정보의 신뢰도 값과 절대적 신뢰도 값인 1의 차이에 대해 자연로그를 취한 값과 비교되었다. 이러한 방법을 사용함으로써 사용자가 지정할 수 있는 임계값은 0에서 1사이의 값이 아닌 0부터 무한대의 값을 가질 수 있게 되었다.

3. 실험 및 실험 결과

3.1 실험 구성

본 논문에서 제안한 확률 기반 필터링 시스템의 성능을 평가하기 위하여 BioNLP Shared Task 2011의 GENIA Task 코퍼스(corpus)를 사용한 실험을 수행하였다. 이 코퍼스는 TEES에서 추출하는 사건 정보들과 같은 사건 유형 및 인수의 역할 유형을 가지는 사건 정보들을 생물학 문헌에 주석처리(annotation)해 둔 코퍼스이다.

먼저 TEES를 사용하여 GENIA Task 코퍼스 중 생물학 문헌의 초록들로부터 생물학 사건 정보를 자동으로 추출하였다. 이렇게 추출된 각 사건 정보에서 인수로 등장하는 유전자(gene) 및 단백질(protein)의 이름은 Hugo Gene Nomenclature Committee (HGNC)에[16] 의해 제공되는 유전자 이름 사전을 활용하여 HGNC의 공식 유전자 심볼로 대체되었다. 이는 동일한 유전자 및 단백질을 서로 다른 이름으로 지칭하는 경우를 파악하고 이를 통일하기 위한 방법이다. 만일 사건 정보의 인수로 등장하는 유전자 또는 단백질 이름이 HGNC 사전에 등장하지 않는 경우에는 원래의 유전자 및 단백질 이름을 그대로 사용하였다.

유전자 및 단백질 이름들이 HGNC의 공식 심볼로 대체된 사건 정보들은 본 논문에서 제시한 확률 기반 필터링 알고리즘을 통하여 참긍정인 것으로 예상되는 것들과 양성 오류인 것으로 예상되는 것들로 나뉘어졌다. 표 2는 실험을 위해 사용된 TEES의 세분화된 정밀도 값으로, TEES의 GENIA Task 코퍼스에 대한 사건 정보 추출 결과를 GENIA Task 코퍼스의 주석 처리된 정보와 비교하여 직접 측정한 값이다. 본 논문의 실험을 위한 정밀도 값은 각 사건 정보의 유형 별로 측정된 값으로, 하나의 사건 정보 유형에 대하여 각 인수의 역할 유형에 따른 정밀도 값은 모두 같은 값으로 책정되었다.

표 2. TEES의 정밀도값
Table 2. Precisions of TEES

	Argument Type			
Event Type	Theme	Cause	Site	Loc
Gene_expression	77.07	N/A	N/A	N/A
Transcription	72.73	N/A	N/A	N/A
Protein_catabolism	80.00	N/A	N/A	N/A
Phosphorylation	71.88	N/A	71.88	N/A
Localization	83.93	N/A	N/A	0.00
Binding	66.67	N/A	66.67	N/A
Regulation	53.21	53.21	53.21	N/A
Positive_regulation	56.88	56.88	56.88	N/A
Negative_regulation	50.79	50.79	50.79	N/A

본 논문에서는 제안하는 필터링 방법의 성능을 횟수 기반(occurrence-based)의 기준(baseline) 필터링 방법과 비교하여 평가하였다. 횟수 기반 기준 필터링 방법은 사건 정보 추출 시스템에 의해 얻어진 사건 정보들 중에서 임계값으로 정한 횟수 이상의 추출 결과들이 나타난 사건 정보들만을 참긍정인 것으로 간주하는 방법이다. TEES에 의해 얻어진 추출 결과에 확률 기반 필터링 알고리즘 및 횟수 기반의 기준 필터링 방법을 각각 적용한 후, 이들 필터링 방법을 통해 참긍정인 것으로 예상된 사건 정보들 중에서 실제로 GENIA Task 코퍼스에 주석 처리 되어 있는 정보들을 실제 참긍정인 것으로 파악하고, 그렇지 않은 것들은 양성오류인 것으로 판

단하였다. 그리고 이를 바탕으로 각 필터링 방법의 양성예측도(true positive rate) 및 위양성율(false positive rate)을 측정하였다. 또한, 각 필터링 방법에 다양한 임계값을 적용하면서 성능 변화를 관찰하였다.

3.3 실험 결과

그림3은 확률 기반 필터링 방법과 횟수 기반 기준 필터링 방법의 임계값을 각각 0에서 10으로 1씩 증가시킨 경우, 각 방법을 통해 참긍정인 것으로 예측된 사건 정보 중에서 실제로 참긍정인 사건의 개수 및 실제로는 양성오류로 판별된 사건의 개수의 변화 추이를 보여준다. 횟수 기반 기준 필터링 방법의 경우 임계값이 어느 수준 이상으로 높아지면 임계값에 따른 성능의 변화가 거의 없어지는데 비해, 확률 기반 필터링 방법을 통해서는 임계값의 변화가 좀 더 지속적으로 필터링 성능에 영향을 미침을 알 수 있다. 또한 횟수 기반 기준 필터링 방법의 임계값으로는 0 및 양의 정수만을 지정할 수 있지만, 확률 기반 필터링 방법을 위해서는 0이상의 모든 실수를 임계값으로 지정할 수 있다. 이러한 점으로 미루어 볼 때, 확률 기반 필터링 방법을 통해 횟수 기반 기준 필터링 방법보다 좀 더 세밀한 필터링 성능의 조절이 가능하다.

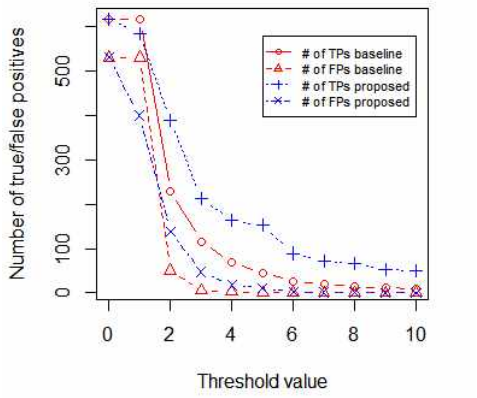


그림 3. 제안된 필터링 방법 및 기준 필터링 방법의 참긍정 및 양성오류 개수의 변화 추이
Fig. 3. Changes of the number of true positives and false positives by proposed and baseline filtering methods

그림 4는 확률 기반 필터링 방법 및 횟수 기반 필터링 방법각각의 ROC 곡선이다. 그림 4를 통해 확률 기반 필터링 방법을 통해 횟수 기반 필터링 방법보다 좀 더 높은 확률로 참긍정인 사건 정보와 긍정 오류인 사건 정보를 구별해 낼 수 있음을 확인할 수 있다.

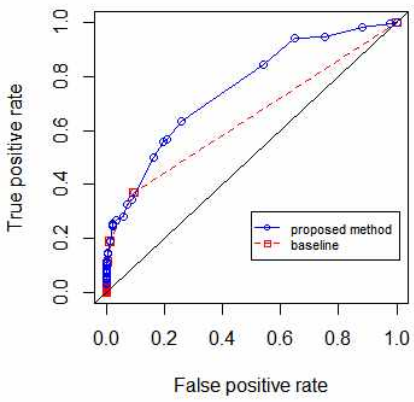


그림 4. 제안된 필터링 방법 및 기준 필터링 방법의 ROC 곡선
Fig. 4. The ROC curves of proposed and baseline filtering methods

표 3은 필터링을 사용하지 않은 경우 및 각각의 필터링 방법을 사용한 경우, 추출된 정보들을 기준으로 계산한 최종 결과의 정밀도, 재현율 및 F값을 표시한 것이다. 특히 각각의 필터링 방법에 대해서는 0을 제외한 여러 임계값들을 사용해 본 결과 얻어진 가장 높은 F값 및 이때 사용된 임계값을 표시하였다. 표 3을 통해, 적절한 임계값이 사용된 경우 확률 기반 필터링을 이용해 정밀도와 재현율의 측면에서 사건 정보 추출 시스템의 성능을 높일 수 있음을 알 수 있다. 이는 정밀도에 좀 더 가중치를 두는 F0.5 값을 통해 더욱 잘 드러난다. 또한 횟수 기반의 기준 필터링 방법이 적용된 경우는 항상 사건 정보 추출 시스템의 전체적인 성능이 저하됨을 알 수 있다.

표 3. 필터링 방법에 따른 정밀도, 재현율 및 F값
Table 3. Precision, Recall and F-values by each filtering method

Method	Threshold	P	R	F1	F0.5
No filtering	N/A	0.51	0.70	0.52	0.53
Proposed method	1.25	0.63	0.48	0.55	0.59
Baseline	1	0.51	0.70	0.52	0.53

확률 기반의 필터링 방법 및 횟수 기반의 기준 필터링 방법은 모두 동일한 정보에 관한 추출 결과가 반복해서 추출될 때, 해당 정보의 신뢰도를 높이 평가하는 방법이다. 따라서, 동일한 정보를 나타내는 추출 결과의 개수를 정확하게 파악할수록 필터링 결과가 정확해 질 것으로 기대되는데, 이를 위해서는 동일한 유전자 및 단백질을 나타내는 서로 다른 유전자

및 단백질 이름들을 잘 파악하는 것이 필요하다. 본 논문에서 이러한 유전자 및 단백질 이름의 동의어 파악을 위해 사용한 HGNC 유전자 이름 사전에는 GENIA Task 코퍼스에 주석 처리된 전체 유전자 및 단백질 이름 중 71%(394/553)가 등록되어 있다. 따라서 유전자 및 단백질 이름의 동의어 파악을 위한 좀 더 효과적인 방법을 사용한다면 필터링 방법의 성능을 더 높일 수 있을 것으로 기대된다.

반면에 본 논문의 실험에서 사용한 정밀도 값은 필터링의 대상이 된 추출 결과들을 코퍼스의 주석 정보와 비교하여 얻은 매우 정확한 수치이다. 이러한 정확한 정밀도 값은 정보 추출 시스템들이 생물학 지식 자동 추론 시스템을 위해 사용되는 실제 환경에서는 얻기 힘든 값이다. 따라서 이러한 실제 환경에서의 필터링 성능은 위의 실험에서 보고한 것보다 저하되어질 수 있을 것으로 보인다.

IV. 결 론

본 논문에서는 텍스트 마이닝을 기반으로 하는 생물학 지식 자동 추론 시스템의 정확성 및 효용성을 높이기 위해 텍스트 마이닝 시스템을 통해 추출된 정보 중에서 긍정 오류일 가능성이 높은 정보를 확률 기반의 방법으로 필터링하는 방법을 제안하였다. 본 논문에서 제안한 방법은 횟수 기반의 기준 필터링 방법보다 높은 성능을 보였다. 또한 적절한 임계값을 적용할 경우에는 전체 텍스트 마이닝의 성능을 높일 수도 있음을 보였다.

본 논문에서 제안한 확률 기반 필터링 알고리즘은 사건 정보 추출 시스템에 특화되어 있다. 그러나 본 논문에서 제안한 추출 결과 및 추출된 정보의 신뢰도 계산법은 다른 형태의 정보를 추출하는 텍스트 마이닝 시스템의 결과를 정제하기 위해서도 긴요하게 사용될 수 있을 것으로 기대된다.

향후에는 추출 결과 및 추출된 정보의 신뢰도를 산정하는 데 있어 개별 사건 정보 추출 시스템으로부터 제공되는 확신값(confidence value)을 활용하는 방안을 연구하고, 효과적인 필터링을 수행하기 위한 임계값을 선정하는 방안에 관해 연구하고자 한다. 또한 이러한 필터링 방법을 적용한 생물학 지식 발견 시스템을 기존의 방법론과[17,18] 연계하여 암 등 복잡질환의 발병 기전을 파악하는데 사용하고자 한다.

참고문헌

[1] P.Zweigenbaum and D.Demner-Fushman, Advanced literature-mining tools, In J.E.Stajich, D.Edwards and D.Hansen, eitors, "Bioinformatics: Tools and Applications," pp.347-381, Springer, Sep. 2009.

[2] E.Antezana, M.Kuiper, and V.Mironovm, "Biological knowledge management: the emerging role of the semantic web technologies," Briefings in Bioinformatics, Vol. 10, No. 4, pp.392-407, May 2009.

[3] T.Slater, C.Bouton, and E.S.Huang, "Beyond data integration," Drug Discovery Today, Vol. 13, No. 1314, pp.584 - 589, March 2008.

[4] Q.Zhu, Y.Sun, S.Challa, Y.Ding, M.Lajiness, and D.Wild, "Semantic inference using chemogenomics data for drug discovery," BMC Bioinformatics, Vol. 12, No. 1, pp.256, June 2011.

[5] C.B.Giles and J.D.Wren, "Large scale directional relationship extraction and resolution," BMC Bioinformatics, Vol. 9, No. suppl 9, pp.S11, Aug. 2008.

[6] D.R.Swanson, "Two medical literatures that are logically but not bibliographically connected," Journal of the American Society for Information Science, Vol. 38, No. 4, pp.228-233, July 1987.

[7] D.R.Swanson, "Complementary structures in disjoint science literatures," In Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, Oct. 1991.

[8] D.R.Swanson, and N.R.Smalheiser, "An interactive system for finding complementary literatures: a stimulus to scientific discovery," Artif. Intell., Vol. 91, No. 2, pp.183 - -203, April 1997.

[9] K.Seiki and J.Mostafa, "Discovering implicit associations between gens and hereditary diseases," In Proceedings of the Pacific Symposium on Biocomputing 2007, Jan. 2007.

[10] M.Yetisgen-Yildiz and W.Pratt, "Using statistical and knowledge based approaches for literature based discovery," Journal of Biomedical Informatics, Vol. 39, No. 6, pp.600 - 611, Jan. 2006.

[11] D.Hristovski, C.Friedman, T.C.Rindflesch, and B.Peterlin, "Exploiting semantic relations for literature based discovery," In AMIA Annual Symposium Proceedings, Nov. 2006.

[12] L.Tari, S.Anwar, S.Liang, J.Cai, and C.Baral, "Discovering drug drug interactions: a text mining and reasoning approach based on properties of drug metabolism," Bioinformatics, Vol. 26, No. 18, pp.i547-i553, Sep. 2010.

[13] J.D.Kim, S.Kraines, W.Guo, and J.Tsujii. "Inference for bioie: Genia meets ekoss," In Proceedings of the 3rd International Symposium on Language in Biology and Medicine, Nov. 2009.

[14] H.J.Lee and J.C.Park, "Towards Knowledge Discovery through Automatic Inference with Text Mining in Biology and Medicine," In Proceedings of the 3rd International Symposium on Semantic Mining in Biomedicine, Sep. 2008.

[15] J.Björne, F.Ginter, J.Heimonen, A.Airola, T.Pahikkala and T.Salakoski, "Extracting Complex Biological Events with Rich Graph-Based Features Sets," In Proceedings of the BioNLP'09 Shared Task on Event Extraction, pp.10-18, June 2009.

[16] A.Cimatti et al., "NuSMV 2: An opensource tool for symbolic model checking," In Proceedings of CAV 2002, pp.27-31. July 2002.

[15] J.D. Kim, S.Pyysalo, T.Ohta, R.Bossy, N.Nguyen and J.Tsujii, "Overview of BioNLP Shared Task 2011," In Proceedings of BioNLP Shared Task 2011 Workshop, pp. 1-6, June 2011.

[16] S.Povey, R.Lovering, E.Bruford, M.Wright, M.Lush and He.Wain, "The HUGO Gene Nomenclature Committee (HGNC)," Human Genetics Vol. 109, No. 6, pp.678-680, Oct. 2001.

[17] S.Leem, K.Wee, "Prediction of SNP interactions in complex diseases with mutual information and boolean algebra," Journal of The Korea Society of Computer and Information, Vol.15, No.11, pp.215-224, Nov. 2010.

[18] H.Jeong, Y.Yoon, "Class prediction of an independent sample using a set of gene modules consisting of gene-pairs which were condition(Tumor, Normal) specific," Journal of The Korea Society of Computer and Information, Vol.15, No.12, pp.197-207, Dec. 2010.

저 자 소 개

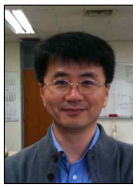
이 희 진

2005 : 카이스트 전산학과 공학사.

현 재 : 카이스트 전산학과 박사과정

관심분야 : 자연언어처리, 바이오인포매틱스

Email : heejin@nlp.kaist.ac.kr



박 종 철

1984 : 서울대학교 컴퓨터공학과 공학사.

1986 : 서울대학교 컴퓨터공학과 공학석사.

1996: Univ. of Pennsylvania Computer & Information Science, 공학박사

현 재 : 카이스트 전산학과 부교수

관심분야 : 자연언어처리, 계산언어학

Email : park@nlp.kaist.ac.kr