KAIST Grand Challenge 30 project
-Finding Research Hypothesis in Natural Sciences-

Aviv Segev

1차년도 연구결과 보고서

Ⅰ 연구 내용 요약

□ 기존 연구 문헌 조사

○ 문제 상황

- 생물학 분야의 방대한 양의 간행물은 이 연구 분야에서 새로운 통찰력과 참신한 가설을 제안하는데 중요하게 활용 될 수 있음.
- 하지만, 이러한 Literature-based 발견 방식은 엄청난 양의 출판물들로 인하여 정보 탐색의 병목현상을 발생시킬 수 있음.

○ 기존 해결책 [1, 2]

- ABC Principle : 연구 간행물에서 언급되는 N_1과 N_2 사이의 V_1의 관계가 있음을 파악하고 N_2와 N_3 사이에 V_2관계가 있음을 알 수 있다면, N_1과 N_3 사이에 V_3의 관계를 추론하는 방식

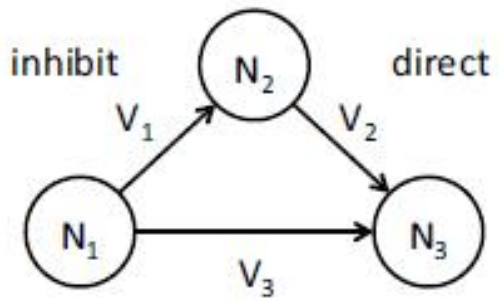

그림 1 ABC Principle

- 따라서, Entity들 (N_1, N_2, N_3) 사이에 관계 (Relationship)을 잘 찾아낼 수 있도록 하는 것이 주요 관건임.
- 본 관계 추론을 위하여 제안된 방법에는 Co-occurrence-based 방법과 NLP-based 방법이 있음.

○ 기존 해결책의 한계 [2]

- Co-occurrence-based 방법은 연구 간행물에서 언급된 횟수를 통하여 관계를 추론하는 방식

한계점 : 언어의 다양한 표현이 정규화되는 과정에서 상실이 일어날 수 있으며, 약간 언급이 되는 관계들에 대해서도 그 중요도의 손실이 발생 할 수 있음.

- NLP-based 방법은 기 존재하는 지식들을 기반으로 하여 그 관계들을 추론하는 방식

한계점 : 우리가 사용할 수 있는 지식들을 생성하는 과정이 필요하며, 지식에 포함되지 않는 관계들을 추론해내기는 어려움.

□ 본 연구팀의 제안 방법

○ Word-embedding 방법 [3, 4]

- 정의 : 자연언어처리의 언어 모델링 방법 중 특정 단어들이나 구들의 언어를 실제 숫자들의 벡터로 구성하는 방법
- 특징 : 단어들의 벡터 연산을 통하여 유사도를 측정하기 쉬우며 벡터 연산을 통하여 새로운 단어들과의 관계를 추론할 수 있음. 현재까지는 뉴스 Corpus를 통한 일상용어의 Vector화 연구가 .

  예) Vector(**King**) - Vector(**Man**) + Vector(**Woman**) ≅ Vector(**Queen**)

□ 1차 년도 결과

○ 연구 수행 방법

- 본 연구팀은 PubMed에서 cancer와 관련이 있는 1500건의 논문 초록을 통해 단어의 말뭉치 (Corpus)를 구성하였음.

○ 연구 수행 결과

- 아래 테이블에 정리되어 있듯이 주요 Keyword들에 대하여 유사 컨셉 단어들을 Biomedical 분야 에서도 찾아낼 수 있음을 확인 함.

Given Concept	10 Nearest Concepts
chemotherapy	therapy, chemo-therapy, recurrent, neoadjuvant, patient, diagnosis, Intra-arterial, FOLFIRI acquire, poor
Cell	apoptosis, effect, human,expression proliferation, growth, cancer, tumor, line
Breast	prostate, lung, colorectal, ovarian, pancreatic, gastric, bladder, mammary
Leukemia	myelocytic, myelogenous, lymphoblastic, lymphotrophic, aml, virus-1-infected, 1,3-bis(2-chloroethyl)-1-nitrosourea, fresh, myeloblastic, leukemic
Benign	high-grade, nonmalignant, superficial, cyst, cirrhosis, micrometastatic

표 1. Associations between Biomedical Concepts

Ⅱ 2차년도 연구 진행 방향

□ 향후 진행 계획

○ 제안된 Word-embedding 방식 확장

- 제안된 방식으로 도출된 결과의 정확성 및 그 활용성을 높이는 방안을 중점으로 연구를 진행할 예정.

○ 추가 연구 방향 모색

- 현재, Biomedical 분야에 국한되어 적용된 방식을 다른 분야 (예, 신소재공학, 물리학 등)에도 적용하는 방안을 연구 중점사항으로도 둘 예정.

Ⅲ 참고문헌

[1] Fleuren, Wilco WM, and Wynand Alkema. “Application of text mining in the biomedical domain.“ Methods 74 (2015): 97-106.

[2] Seki, Kazuhiro, and Kuniaki Uehara. “Supervised Hypothesis Discovery Using Syllogistic Patterns in the Biomedical Literature.“ IJCAI. 2013.

[3] Mikolov, Tomas, et al. “Efficient estimation of word representations in vector space.“ arXiv preprint arXiv:1301.3781(2013).

[4] Mikolov, Tomas, et al. “Distributed representations of words and phrases and their compositionality.“ Advances in neural information processing systems. 2013.