글 + 사진. 김동욱 | 객원기자

생물학과 전산학이 만나는 곳, KAIST NLP&CL 연구실

암 관련 유전자 검색엔진, 온코서치

암은 수천 개 이상 되는 유전자의 비정상적 변화와 이에 따른 신호 전달 체계의 교란으로 발생한다. 암에 관계된 유전자의 수가 워낙 많고 관련 연구도 방대하기에 꼭 맞는 관련 문헌을 찾는 일은 쉽지 않다. '온코서치'는 수백만 건의 논문에서 암 관련 유전자에 대한 논문을 자동으로 추출하여 검색할 수 있는 검색엔진이다.

구글의 공동 창업자 세르게이 브린과 래리 페이지는 '페이지 랭크(page rank)'라는 간단한 개념으로 네트워크 상의 정보를 체계화하는 방법을 구현했다. 사용자는 인터넷에서 기존의 검색엔진보다 더 쉽고 정확하게 자신이 원하는 정보를 찾을 수 있었고, 이는 네트워크 상에서의 정보 교환을 촉진했다. 결과적으로 네트워크 상에 이전보다 유용한 정보의 양이 많이 누적되는 데 촉매제로 작용했다. 암과 유전자 간의 관련성을 검색할 수 있게 하는 '온코서치(OncoSearch)'의 지향점도 바로 거기에 있다. 암과 유전자의 관계를 다룬 수많은 논문 중에서 자신의 연구와 관련된 논문을 더 쉽고 정확하게 체계화하여 검색할 수 있도록 하겠다는 것이다. 구글이 그랬던 것처럼 온코서치도 암을 연구하는 연구자들의 정보 교환을 촉진하는 촉매제의 역할을 하게 될 것이다.

Oncology + Search = OncoSearch

온코서치는 종양학을 뜻하는 '온콜로지(oncology)'와 검색을 뜻하는 '서치(search)'를 합성해서 만들어진 이름이다. 말 그대로, 종양학과 관련된 논문을 검색할 수 있는 검색엔진이다. 보다 자세히 설명하면, 온코서치는 수많은 종양학 논문 중에 암과 유전자의 관계를 기술하는 문장을 분석하여 둘의 관계를 자동으로 추론한다. 예를 들어, "A 유전자(gene)는 종양억제유전자(tumor suppressor gene)다"라는 문장과 "A 유전자의 발현을 촉진(overexpression)했을 때 암 세포가 죽고 성장이 퇴보(regression)되었다"라는 문장이 있다고 하자. 후자는 그 강도가 약할 뿐, 전자와 유사한 의미를 갖는다. 실제 대부분의 논문에서는 후자처럼 묵시적으로 표현한다. 이 때문에 키워드 중심으로 논문을 검색하면 의미 있는 연구 결과를 찾지 못하는 경우가 빈번하게 발생한다. 온코서치는 후자처럼 표현된 문장의 의미를 추론하여 전자의 문장처럼 이해한다. 연구자는 유전자와 암의 종류를 키워드로 입력하고, 검색하고자 하는 종류만 선택하면 관련 문헌을 쉽고 빠르게 검색할 수 있다. 문장에 표현된 암과 유전자와의 관련성은 유전자 변화 양상(이하 CGE: Change in Gene Expression), 암 변화 양상(이하 CCS: Change in Cell State) 및 둘 사이의 인과관계(이하 PT: Proposition Type) 정도를 수치화하여 측정한다. 약 300만 건의 생물학, 의학 논문이 저장된 메드라인(Medline)에서 논문을 읽어오고, 각 논문의 초록을 문장 단위로 분해한 후 유전자와 암에 관련된 키워드를 기반으로 각각의 문장을 구조화한다. CGE는 사건 정보 추출 시스템을 이용하여 측정하고, CCS 및 PT는 800여 개의 문장으로 학습한 최대 엔트로피 분류기를 사용하여 측정한다. 이렇게 수치화한 CGE, CCS, PT 값을 바탕으로 발암유전자(oncogene), 종양억제유전자, 생체지표(biomarker)인지 분류하여 검색 유형에 맞는 값만을 사용자에게 제시한다. 세 가지 값에서 도출한 신뢰도 점수와 확률값의 조화평균(harmonic mean)을 찾아 높은 값의 문장을 갖고 있는 논문부터 사용자에게 보여준다.

Biology + NLP = BioNLP

온코서치는 전산학에 기반한 연구이지만, '핵산 연구(Nucleic Acids Research)'라는 생물학 학술지에 논문이 실렸다. 온코서치를 개발한 NLP & CL(Natural Language Processing and Computational Linguistics) 연구실의 박종철 교수는 자신의 연구실을 '자연언어정보서비스'를 하는 곳이라고 정의했다. 자연언어처리 분야의 기술이 있는데, 이 기술을 어떤 가치가 있는 곳에 적용할 것인가를 고민한 후 연구를 시작했다고 한다. 글로 쓴 문장을 수화 애니메이션으로 변환하는 연구, 글이나 대본에서 감정을 추론하는 연구나 치매와 관련된 연구도 그렇게 시작되었다. 기술의 응용을 고민하다 보니 다양한 분야와 융합하게 되고 타 분야의 학술지에 실리는 경우도 종종 생긴다. 온코서치는 전산학의 분야인 자연언어처리와 생물학이 접목한 바이오 자연언어처리(이하 BioNLP)에 속한다고 할 수 있다.

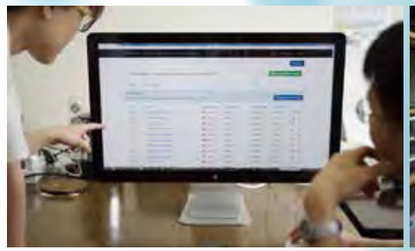
BioNLP를 시작하게 된 계기를 물었다. "제가 처음에 BioNLP을 시작한 것이 2001년이었어요. 대학원에서 같이 공부했던 친구 중에 졸업하고 싱가포르 대학에서 교수를 하는 친구가 있는데, 이 친구가 단백질 간의 관계가 적힌 64개의 문장을 주면서 '혹시 자연언어처리를 통해 단백질 간의 관계를 자동으로 추출할 수 있겠느냐?'라고 했던 것이 계기가 되었죠."

온코서치는 NLP & CL 연구실에서 진행하는 BioNLP 연구의 연장선상에 있다. 미래창조과학부에서 중견연구자과제 공고가 떴을 때 광주과학기술원(GIST) 이현주 교수와 같이 암에 초점을 맞춘 연구를 기획하면서 시작되었다. 온코서치 연구를 주도적으로 진행한 이희진 박사과정생은 융합학문인 BioNLP를 전공하는 것에 대해서 어떻게 생각했을까? "위험부담은 있었지만 취지가 좋더라고요. 생물 교과서도 보고 다 공부했죠. 근데 힘든 건 생물학과 전산학은 사고와 패러다임이 다르다는 거였어요. 처음에는 되게 당황했어요. 조금씩 얘기하다가 서로를 이해하게 된 거죠. 융합학문을 할 땐 서로 열린 마음을 갖는 것이 필요한 것 같아요. 나만의 패러다임을 계속 요구하는 것은 좋지 않은 것 같아요."

연구실에서 진행되는 후속 연구는 어떠한 것이 있는지 물었다. 이희진 박사과정생은 졸업 후에도 온코서치가 수집한 대량의 정보를 종합해서 새로운 정보를 추론하는 시스템을 연구할 것이라고 한다. 박종철 교수는 치매에 관련된 연구를 언급했다. 개인의 일상적인 언어 습관을 분석하여 언어 구사가 반복적으로 되고 뭉개지는 시점을 미리 포착하여 치매로 발전하기 전에 알려주겠다는 것이다.

끈기, 자신감, 패기를 가져라

융합학문을 하고자 하는 후학들에게 하고 싶은 말이 있는지 물었다. 박종철 교수는 끈기, 자신감 그리고 패기의 세 가지를 가져야 한다고 강조했다. "내게 중요한 것은 다른 사람에게도 중요한 것이라는 자신감이 중요해요. 이러한 자신감에 기반한 끈기, 인내력도 필요하죠. 융합이 공부해야 할 것도 많고 결과도 쉽게 나오지 않으니까, 지치지 않는 것이 중요해요. 마지막으로, 패기 있는 친구들이 많이 왔으면 좋겠어요. 가치 있다고 생각하는 연구를 직접 해보겠다는 생각을 가진 학생이 더 많아졌으면 좋겠어요. 나 아니면 누구도 할 수 없다는 마음 같은 거죠."



KAIST NLP & CL 연구실
박종철 교수