**인문학적 자원을 활용한 디지털 큐레이션의 가능성에 대한 토론문**

김광림 (네이버 머신러닝 엔지니어)

도원영 선생님의 발표문에서 많은 것을 배웠습니다. 저는 사전학과는 거리가 멀지만 디지털화와 데이터 구축 및 활용에서는 가까운 입장이기에, 디지털 환경에서 사전 이용자의 관점과 데이터과학 및 AI연구 관점을 아울러 의견을 개진해보겠습니다.

선생님께서 사전 이용자는 수많은 사전과 정보가 있음에도 불구하고, ‘내가 원하는 사전은 없다’는 처지라고 말씀하셨습니다. 연구자든 일반인이든 가릴 것 없이, 늘 겪는 문제를 제대로 짚어 내셨습니다. 저는 여기서 조금 구분이 필요하다고 생각합니다. 학술 연구자가 아닌 이상, 일반인은 ‘사전’이 있는지를 따지기보다 정보가 사전에 없다고만 생각합니다. 사전에 정보가 없다면 사용자는 다른 대안을 선택하게 되는데요, 보통 검색엔진에 질의를 던져 관련된 문서를 찾아보기 마련입니다. 정보를 찾는 사용자의 행태를 볼 때, 선생님께서 제시한 해결방안 중 ‘어휘지식정보가 있는 곳으로 안내‘에서 분기되는 새로운 해결책도 고려하실 수 있지 않을까요? ‘사용자의 의도를 파악해 가장 근접한 정보부터 찾아갈 수 있는 방법‘이 바로 그것인데요. 이미 ‘네이버 지식백과’ 등 온라인 사전에서 유사한 표제어나 연관 표제어를 제시하고 있습니다. 표제어간의 언어적 연결성을 넘어 의미와 내용상의 관련성을 데이터로 구축하여 ‘한 차원 높은 근접 정보’를 사전에서 보여주는 것이 필요하다고 생각합니다. 선생님께서 소개하신 ‘사전의 지도’는 이를 위한 기본 자료로 적합한 마중물이 될 수 있으리라 생각합니다.

한국어 은유 의미망 연구는 디지털 데이터를 실용적으로 이용하는 현업의 입장에서 매우 유용하다고 생각합니다. 전문가의 사전은 이래야 한다는 것을 잘 보여주고 있습니다. 은유와 비유는 언어마다 다르고, 인간 인지의 차원 높은 사고의 결과이기에 그 사례도 많거니와 이것만 따로 정리해 두기 어렵습니다. 그래서 AI연구에서 은유 의미망을 정리한 사전은 가치가 아주 높습니다.

일례로 예시를 드셨던 ‘산 입에 거미줄 치랴’를 네이버 파파고와 구글 번역으로 보자면 다음과 같습니다.

* 네이버 파파고 : A living man's mouth must be covered with cobbled together.
* 구글 번역 : A spider web in the mouth of the living.

위 예시는 속담으로 한영사전에 예문이 다음과 같이 있습니다. “Everyday brings its bread with it.” 그런데도 정확하지 않은 직역을 하고 있습니다. 사전에 있는 문구도 그대로 활용하지 않는다는 점도 확인할 수 있습니다. 하지만 핵심은 이것이 아니지요. 기계번역은 날로 발전함에도 불구하고 위와 같은 결과를 보여줍니다. 네이버에서도 한국어 AI 연구를 하면서 크게 아쉬움을 느끼는 부분이 바로 전문가가 검증한 정확한 데이터가 소수라는 점입니다. 정확한 은유/비유 의미망 데이터가 있다면 번역 품질은 크게 향상될 것입니다. 또한 최근의 AI 연구는 컴퓨터가 정말 ‘이해를 할 수 있는지’, 그리고 ‘의미를 설명할 수 있는지’에 초점에 맞춰져 있습니다. 은유 의미망 연구결과는 최신 AI 개발에서도 중요하게 적용될 수 있다 생각합니다.

사전학 연구에서 이런 부분에 초점을 맞춰 인간, 그 중에서도 ‘전문가’만이 할 수 있는 연구 주제가 나오는 분위기인지 선생님께 여쭙고 싶습니다. AI 연구를 가까이 지켜보는 입장에서, 많은 과제들이 인문학 전문가 없이 상당수 ‘잘’ 진행되는 것을 볼 때 위기감을 느끼고 있습니다. 인문학자들은 사후 ‘검수’와 ‘검증’만 요구받을 수도 있는 처지입니다. 처음부터 협력연구로 더 나은 성과를 보이기 위해서는 적극적으로 AI가 하지 못하는 영역을 찾아야 한다고 생각합니다. 이는 이미 많은 분들이 이야기 해 온 것을 다시 한번 반복하는 것에 지나지 않기에, 실천이 필요한 시점입니다.

마지막으로 선생님께 드리고 싶은 질문은 ‘사전에는 재미있는 이야기가 들어있기 힘든가?’입니다. 원래도 인류에게 이야기는 그 자체로 중요했지만 특히 요즘은 ‘스토리텔링’이란 화두로 어디서든 그 중요성이 지나칠 정도로 강조되고 있습니다. 사전 이용자는 과연 ‘내가 찾는 정보’가 없어서 ‘원하는 사전이 없다’고 생각하는 것일까요? 내가 알고자 헀던 것은 아니지만 연관 있는 재미있는 정보가 있다면 어떻게 반응할까요? 이는 위키피디아와 나무위키의 사용에서 곱씹어 볼 여지가 있습니다. 학생들이 학교 과제용으로는 네이버 지식백과를 보고 적어내지만, 배경지식과 흥미있는 관련 정보를 알려면 나무위키를 본다는 사실은 이미 잘 알고 계신 바입니다. 나무위키에는 이야기가 있습니다. 정확성과는 거리가 먼 내용이 많지만, 지식백과와 나무위키 두 개를 같이 이용하는 학생들은 이를 충분히 알고 나름 활용하고 있습니다. 저는 이 점에서 사전에서도 서로 연관되는 정보를 잘 엮어 ‘스토리텔링’이 들어간 큰 주제 사전을 제공할 가능성이 있지 않을까 생각합니다. 이미 정확한 정보가 있으면 AI가 자연스럽게 문장을 만드는 것은 시간 문제가 되고 있습니다. (GPT-3 가 대표적인 예가 되겠습니다.) 이를 대비해 사전마다, 그리고 사전 안에서 어휘 근접성 수준이 아닌 고차원의 의미 연결망 데이터를 구축한다면 보다 적극적으로 기술 변화에 대처할 수 있을 것이라 생각합니다. 이러한 시도가 이미 학계에서도 고려중일 것이라 생각하기에, 학계 현황과 앞으로의 방향이 궁금합니다.