국어 정보화와 전문용어 표준화의 선구자
— 최기선 한국과학기술원 교수를 만나다

최기선(한국과학기술원 전산학과 교수)

질문자 이경우(서울신문 어문팀장)
때 2016. 6. 8.(수) 곳 교수 연구실

1970년대 중반 대학생이 된 청년은 수학을 택했다. 그런데 그에게 그만 컴퓨터가 눈에 쏙 들어오고 말았다. '컴퓨터가 언어를 이해할 수 있다면? 사고를 한다면?'이라는 생각에까지 미쳤다. 도서관에 살다시피 하면서 관련 서적을 탐독하기 시작했고, 학술대회를 쫓아다녔다. 그러던 어느 날 꼭 읽어야 할 책이 생겼는데, 거금 3만 원이었다. 어머니는 그런 책이 있는지 믿기지 않아 아들과 함께 서점에 갔다. 대학가 한 달 독방 하숙비가 3만 원, 80㎏들이 쌀 한 가마니가 2만 원 정도 할 때였다.

전산학과가 없던 시절, 대학원에서는 응용수학을 할 수밖에 없었지만, 그는 본격적으로 언어공학을 파고들기 시작했다. 언어와 관련한 지식이 필요하다는 생각에 국어학 수업도 열심히 들었다.

국어 정보화, 그는 컴퓨터가 국어를 분석하고 이해하도록 하는 일을 개척해 나갔다. 이것은 국가의 힘을 키우고 지식정보화 시대, 정보에서 소외되는 사람이 없게 하는 것이기도 했다.

주시경 선생은 일찍이 "말이 오르면 나라가 오른다."라고 말했다. 그는 전문용어가 정비돼야 이 분야가, 과학이 발전한다고 여겼다. 또한 사회 수준

이 전반적으로 같이 올라간다고 믿었다. 그래서 전문용어를 표준화하는 일에도 발 벗고 나서게 됐다. 그는 전문용어가 제대로 정비되는 생태계를 만들어야 한다고 말했다.

2015년 한글날 그동안의 공로를 인정받아 옥조근정훈장을 받았다.

이경우 선생님, 반갑습니다. 지난해 한글날상을 받으셨는데 늦었지만 축하드립니다.

최기선 고맙습니다.

이경우 축하를 많이 받으셨을 텐데요. 어떤 소감을 전하셨는지요.

최기선 한글날 수상자들이 대개 어문 계열에 있는 분입니다. 공학계이고 전산학에서 문화 관련 상을 받은 게 뜻깊었습니다. 한글은 휴대전화에서도 쓰고 우리 생활에 밀착돼 있습니다. 여기에다 한국어 음성으로 검색하는 게 공학이니까 거기서 평가받았다는 데 큰 보람을 느낍니다.

이경우 선생님처럼 전산학 분야에서 국어와 관련해 상을 받으신 분이 없는 것으로 알고 있습니다. 국어 정보 처리 분야에서 오랫동안 연구를 해 오셨고 만만치 않은 족적을 남기셨다고 들었습니다. 전산학을 하시면서 국어와 관련해 연구하시는 게 특별해 보이는데, 어떤 계기가 있었는지요.

최기선 대학교 1학년 때 컴퓨터가 막 쓰이기 시작했죠. 학교 전체에 컴퓨터가 한 대 있을 때였죠. 컴퓨터가 막 쓰이기 시작할 무렵인데, 막연하게 컴퓨터가 말을 하거나 생물학적인 구조가 있다면 대단하겠다는 생각을 했죠. 확실치 않지만 어떤 글을 읽었던 것도 같습니다. 이런 생각을 나도 모르게 믿었죠. 그래서 이쪽으로 많이 생각하고 있었습니다. 그러다가 한국과학기술원(이하 '카이스트')

석사 과정에 입학하게 됐습니다. 카이스트가 홍릉에 있을 때였죠. 한국외국어대도 홍릉에 있었는데요, 박사 과정 1학년 어느 날 우연히 거기를 가다가 게시판에서 서울언어학국제학술대회(Seoul International Conference on Linguistics)가 열린다는 안내문을 보게 됐습니다. 그 프로그램을 보니까 컴퓨터를 가지고 언어를 분석하는 게 있더라고요. 잘 몰랐지만 가 봤습니다. 가서 보니까 제가 배우는 오토마타(auto-mata) 이론이 이용되더라고요. 관련된 책을 하나 알게 됐는데, 그 책을 한번 꼭 보자는 마음이 생겼습니다. 그래서 서울 종로 종각에 있는 종로서적센터 외서부에 가니까 그 책이 있더라고요. 너무 비싸더군요. 당시 3만 원이었습니다.

이경우 그 책을 샀습니까.

최기선 네, 그런데 학교에 사 달라고 할 생각을 못 하고 어머니께 사 달라고 했죠. 어머니께서 무슨 그런 책이 있느냐고 하시더군요. 그러면서 직접 같이 가 보자시더니 두말없이 그 책을 사 주시더라고요.

이경우 어머님이 같이 가셨어요?

최기선 너무 믿기지 않는 큰돈이니까 어머니가 가서 직접 보자시더군요. 정말 읽고 싶다고 했죠. 읽고 싶더라고요.

이경우 정말 호기심이 많으셨나 봅니다.

최기선 그랬던 거 같아요. 그 책을 한 열 번쯤 읽으니까 알겠더라고요. 그게 시작이었습니다. 그다음부터는 하루 반은 도서관에 살았습니다. 좀 막무가내로 했죠.

이경우 선생님께서 대학 들어갈 당시만 해도 전산학이라는 것이 낯선 학문이었잖아요. 우리나라에 거의 갓 들어왔을 때였나요.

최기선 그렇죠. 저는 대학에서는 수학과를 나왔습니다.

이경우 아, 대학에서는 원래 수학을 공부하셨군요.

최기선 네, 학부 때는 수학을 공부했습니다. 그런데 옆에 계산통계학과가

생겼어요. 계산통계학의 김현택 교수님께서 컴퓨터를 가르치셨지요. 카이스트 들어올 때도 응용수학과였습니다. 들어와서 1년 지날 때 전산학과라는 게 생겼죠.

이경우 그러니까 처음부터 전산학을 하신 게 아니었군요.

최기선 뭐, 없었으니까요.

이경우 연구실 오기 전에 선생님 공적이 담긴 보도자료를 봤는데요. 봐도 잘 모르는 내용이 있더군요. '국어 자연언어 정보 처리 분석기'를 만드셨다고 돼 있습니다. 얼른 머릿속에 들어오지 않는데 어떤 일을 하신 건지요.

최기선 현재의 정보 검색, 인터넷 검색이 가능하게 한 것입니다. 제가 1992년께 '정보 검색'이라는 강의를 열었습니다. 지금 네이버에서 검색하는 프로그램을 만든 사람들이 그때 학생들이었지요. 지금 컴퓨터에서 한국어를 다루는 데 기본적인 일들을 맨 처음 다 했습니다. 예를 들어 화학공학에는 화학물질을 분석하는 기술이 있습니다. 새로운 화학물질을 만들어 내기도 하는데, 그때 실험 장비가 필요하잖아요.

이경우 그렇겠죠.

최기선 그런 실험 장비라는 게 결국은 말뭉치 또는 코퍼스(Corpus)라는 것이고요. 그것을 만드는 일을 우리나라 최초로 한 거죠. 대량으로 그것을 많이 만들고 공개했습니다. 그걸 바탕으로 '21세기 세종계획'이라든지 이런 것들이 가능하리라 보게 됐습니다. 스펠링을 고치는 프로그램들도 최초로 제가 다 만들었죠. 이런 것을 보고 사람들이 '되는 구나' 하고 따라서 한 거고요.

이경우 좀 더 구체적으로 자연언어 처리 분석기에 대해 설명을 해 주시죠.

최기선 구글이나 네이버에서 검색하는 데 들어가는 프로그램이나 이론들이죠. 그런데 그걸 한국어가 되게, 한글 처리가 되게 기본적인 일

들을 한 거죠. 이것을 최초로 만들어서 이쪽 분야의 바탕이 되게 했습니다.

이경우 이제는 컴퓨터가 한국어를 인식하는 능력이 훨씬 발전하지 않았습니까.

최기선 그렇죠. 옛날에는 키워드를 쳐서 검색을 했지만 지금은 문장을 쳐도 이해할 수 있습니다. 또 요약까지 해 주기도 합니다. 구글에서 '버락 오바마'를 치면 오바마에 대한 특성들이 간추려져 일목요연하게 나오기도 하잖아요. 거기에 필요한 기본적인 일들을 지금도 하고 있습니다.

이경우 그럼 거기에 있는 엄청나게 많은 정보, 즉 빅데이터를 가지고 언어 변화, 언어 분석을 쉽게 해 줄 수 있는 거죠?

최기선 그렇습니다. 빅데이터는 여러 가지 측면이 있는데요, 언어 분석이라고 하는 것을 제가 처음 했을 때는 문법을 직접 썼어요. 규칙을 다 쓰고 문장의 구조를 언어학의 이론대로 만들었습니다. 지금은 확률 통계로 해서 빅데이터에서 반복되는 패턴을 봅니다. 이런 조건일 때는 이렇다든가 하는 경우의 수가 많아지면 예측을 더 정확하게 할 수 있는 거죠. 지금 상업적인 분야에서도 많이 이뤄지고 있습니다.

이경우 지금은 지식정보화 시대이고 정보가 넘쳐 나고 있습니다.

최기선 수많은 웹에서 만들어진 지식정보들을 추출해 내야 하는데요. 텍스트도 굉장히 많죠. 이에 비례해서 지식정보에 대한 것도 점점 커지고 있거든요. '시맨틱 웹'이라는 분야에서도 커지고 있는데, 이것 역시 또 하나의 빅데이터입니다. 지식이라고 하는 것과 웹 텍스트가 커지면 커질수록 상응하는 패턴들이 굉장히 많아져서 앞으로는 텍스트를 지식화할 수 있습니다. 무슨 일을 하는 건지 이해할 수 있는 방법을 기계에 부탁할 수 있는 거지요.

이경우 최근 서울 강남역과 구의역에서 사망한 이를 추모하는 포스트잇을 많이 붙이는 일이 일어났습니다. 이 포스트잇에 들어 있는 글들을 분석해 어떤 현상들을 읽을 수도 있겠습니다.

최기선 그렇죠.

이경우 감정까지 읽을 수 있을까요.

최기선 감정도 일종의 분류 체계라고 보면요. 우울한 것도 있고, 화난 것도, 기쁜 것도 있습니다. 여러 가지 세분화된 잣대들이 있는 거죠. 그래서 나타난 것과 감정의 대응이 결국 하나의 지식 체계를 만들게 되는 것이지요. 이것도 빅데이터화해서 감성 분류를 할 수 있는 거죠.

이경우 요즘 인공지능 얘기를 많이 접하게 됩니다. 지난번 알파고와 이세돌의 바둑을 통해서 더 그렇게 됐지만요. 컴퓨터가 인간의 언어를 배우는 게 아니라 '습득'한다는 말도 나올 수 있는 건가요.

최기선 그렇게 볼 수 있습니다. 습득한다고 하는 것은 머신 리딩인데, 리딩을 한다는 것은 글을 읽고 컴퓨터가 이해를 한 뒤에 어떤 것이 되게 하는 것입니다. 거꾸로 사람한테 자신이 무엇을 알고 있는데 부족한 게 뭐냐, 또는 어떻게 하느냐 하는 것을 얘기할 수도 있습니다. 이런 것을 머신 리딩이라고 하는데, 기계가 독해를 한다는 것이죠.

이경우 이게 지금도 충분히 가능한 건가요.

최기선 가장 뜨거운 분야죠. 지금 학계에서는 열심히 노력하고 있는 분야죠. 이게 언어학적인 이론도 있지만 머신 러닝이나 수학적인 모델에 의해서 어떻게 분석하느냐도 중요한 부분입니다. 이것이 빅데이터이기 때문에 너무 양이 많아서 컴퓨터가 빨리 할 수 있는 방법을 찾아야 하는 것이 큰 문제입니다.

이경우 좀 전에 '시맨틱 웹'이라는 용어를 사용하셨습니다. 선생님의 최고

관심 분야라고 들었습니다.

최기선 자연언어 처리는 문장을 어떻게 이해하고 해석할 것인지에 대한 분석이나, 컴퓨터가 문장을 생성하는 것과 언어 자원, 실험 장치들을 얘기하는 체계입니다. 시맨틱 웹은 그것으로 인해 만들어지는 웹의 다음 세상을 얘기하고 있는 거죠. 지금 웹에서는 웹 페이지를 클릭하면 링크가 있고, '위키피디아'라는 것이 있습니다. 그리고 위키피디아를 정제해 구축한 '디비피디아'라고 하는 데이터베이스들도 있습니다. 시맨틱 웹은 그걸 좀 더 개념화하고 체계화해서 추론도, 이해도 되게 하는 하나의 차세대 웹이죠.

이경우 지금보다 정교해진 다음 버전이라고 생각하면 되나요.

최기선 다음 버전이죠. 그러니까 사람이 이해하는 것만큼 기계가 똑같이 이해할 수 있는 상태인 거죠.

이경우 그럼 그다음은 없는 거겠네요. 사람만큼 이해하게 되는 거니까요.

최기선 그렇죠. 궁극적인 목표는 다 이해를 하게 하는 것이 되겠지요.

이경우 그러면 시맨틱 웹이 궁극적으로 추구하는 목표라고 해야 할까요. 그건 무엇인지요.

최기선 우선 사람이 할 수 있는 것 가운데 이해를 하고 추론을 해서 증명할 수 있는 것들을 기계가 다 할 수 있게 하는 거죠. 예를 들어 우리가 주식 투자를 할 때 현재 상황이 어떤데 어떤 결론에 의해서 하게 되잖아요. 이런 정보를 전부 기계가 이해해서 투자를 해야 할지 말지 의사 결정을 내릴 수 있는 단계가 시맨틱 웹입니다. 그렇지만 기계가 할 수 있는 것과 사람이 할 수 있는 것은 분류가 된다고 봐야죠.

이경우 그렇군요. 저는 이 말씀을 듣고 그럼 사고하는 것조차도 기계가 다 해 주는 것 아닌가 하는 우려가 들기도 했습니다. 그래서 인간이 더 퇴보하는 것 아닌가 하는 의심도 해 보게 되고요.

최기선 그러면 인간은 더 좋은 일을 할 수 있겠죠. 예를 들어 어떤 연구를 할 때 나와 비슷한 연구를 하는 사람이 어디까지 했는지 아는 게 엄청난 일이거든요. 기계가 그것을 어느 정도 알려 주면 저는 그다음 일을 하면 되는 거죠. 연구 속도가 더 빨라지는 겁니다.

이경우 선생님, 그런데 컴퓨터는 어떤 방식으로 대상을 인식하게 되는 건지요.

최기선 지금 시맨틱 웹 세상에서는 위키피디아를 보면 200개 언어가 링크돼 있습니다. 우리나라의 박근혜 대통령은 한글로 쓰인 한국어로 돼 있지만 영어, 독일어, 중국어로도 돼 있지요. 각각의 언어로 된 페이지가 다 있습니다. 그러면 시맨틱 웹은 개념의 세계이니까 이 세상의 모든 사물에 대해 일련번호, 즉 주민등록번호 같은 것을 주게 됩니다. 유아르아이(URI: Uniform Resource Identifier)라고 해서요. 만일 박근혜 대통령이다 하면 웹에서 볼 수 있는 유아르아이가 있습니다. 한글로 어떻게 쓰고 영문자로 어떻게 쓰는지 미국에서는 어떤 방식으로 기술하고 있는지 다 다르겠죠. 하지만 사람에 대해서는 어떤 말을 써도 하나인데, 여러 가지 말로 번역돼 있거든요. 이런 것들이 시맨틱 웹까지는 안 가지만 다국어 데이터 웹이라는 것으로 하나의 간접적인 번역이 되는 거죠.

이경우 언젠가는 컴퓨터가 인간이 하는 모든 말을 이해할 수 있겠네요.

최기선 궁극적으로 그렇게 되는 거죠.

이경우 외국어를 배우는 게 아주 어려운 일인데, 컴퓨터는 쉽게 배워서 사람보다 통역을 잘하겠습니다.

최기선 그걸 지향하고 있죠. 학술적으로나 상업적인 분야에서도 그렇고요. 빅데이터 문제니까 데이터가 쌓이면 쌓일수록 더 잘하게 되는 거죠. 한국어가 다른 나라 언어에 비해 얼마나 데이터화가 많이 돼 있느냐가 중요한 문제입니다. 위키피디아나 웹에 있는 한국어

백과사전들 크기가 영어의 10분의 1도 안 되지요.

이경우 그 정도인가요.

최기선 네, 영어는 약 500만 페이지인데 한국어는 35만 페이지 정도밖에 없으니까요. 그러니까 우리나라 말이 더 많아질수록 더 쉽게 다른 나라말하고도 소통할 수 있게 됩니다. 그 상태가 되면 우리나라의 문화적인 힘도 커지는 거지요.

이경우 구글이 한국어를 번역하는 데 오류를 많이 낸다고 하는데 이것은 한국어와 관련한 데이터가 적어서 그렇다고 봐야 하나요.

최기선 적은 거죠. 직접적인 텍스트 양도 적고 웹에 떠 있는 백과사전의 양도 엄청나게 적습니다. 그리고 아까 말씀드렸듯이 화학공학에서 실험을 하려고 하면 실험 장치가 있어야 하잖아요. 거기에 해당하는 사전 같은 것들도 아주 적죠.

이경우 일부에서는 한국어가 다른 언어에 비해 구조적으로 더 복잡하고 어려운 면이 있어서 그런 것이라고 말씀하시는 분도 있습니다.

최기선 제 생각에는 그렇지는 않은 것 같습니다. 영어는 수요가 많고 실험 장치를 만드는 사람도 많고 실험 장치도 굉장히 다양합니다. 한국어는 그렇지 못하다는 것뿐이죠. 영어와 관련해서는 미국에서 공용 데이터베이스를 구축해 놨어요. 펜실베이니아 대학 등에서 해 놓은 것도 있고 굉장히 많습니다. ≪월스트리트저널≫ 같은 곳에서도 연구하라고 자기네 신문 데이터베이스를 통째로 다 줬습니다. 라이선스 문제 같은 게 하나도 없습니다. 사람들이 단어에 대한 분류도 해 주고, 의미도 데이터베이스에 다 올려놓습니다. 아까 말씀드린 유아르아이가 뭐다 등 온갖 것을 뱅크에 다 올려놓습니다. 많으니까 '뱅크'라고 부르죠. 그것으로 연구를 합니다. 우리나라에는 없거든요. 우리는 거기의 밑바닥에 좀 깔린 상태라고 봐야죠.

이경우 우리나라는 그동안 별 관심이 없었던 건가요.

최기선 관심은 있죠. 영어의 경우에는 그것을 쓰려는 사람이 많고 그 체계에다 모든 것을 공개하는 원칙이 있습니다. 돈을 받긴 받아요. 돈을 안 받는 데도 물론 많고요. 돈을 주면 장치들을 사서 연구를 할 수 있는 모든 장치가 한꺼번에 있는 거죠. 우리나라는 그것을 좀 하긴 했는데 여러 가지 저작권법도 있고 돈을 주고도 살 수 없는 것들이 많은 거죠. 숨겨진 자료들이라고 할 수 있습니다. 우리나라의 연구 평가 제도가 기술료를 받아야지 평가를 받는 체계여서 그렇죠. 이것을 남한테 공개해서 얼마나 쓰였는가에 따라 평가를 하면 안 그럴 텐데, 기술료를 얼마나 많이 받아서 수입이 생겼느냐에 따라 평가를 하거든요. 그러다 보니까 공유에 대한 미덕을 잘 모르는 거예요. 그것 때문에 발전을 못 하는 거죠. 그래서 저변 확대가 안 되는 것이고요.

이경우 정확하게는 언어공학이라는 말을 사용하는데 이 분야와 국어의 발전은 어떤 관계가 있는 건지요.

최기선 서로 시너지 효과가 분명히 있죠. 제가 박사 과정 때 연세대에 가서 국립국어원장을 지내신 남기심 교수님과 영어영문학과 이익환 교수님의 강의를 한 학기 들었습니다. 제가 이쪽 분야에서 한 로직 프로그램으로 규칙을 어떻게 쓰는가에 대해 그쪽에서 받아서 하시는 분들도 계시죠. 지금은 언어공학의 방법으로 국어사전에 용례를 많이 찾아서 싣고 있습니다. 그리고 사람들이 많이 쓰는 정의에 대한 태그도 만들어 줄 수 있습니다. 그러면 국어사전들을 좀 더 생활 밀착형으로 만들 수 있는 거죠.

이경우 사전 말씀을 하시니까 아까 잊었던 게 생각나는군요. 선생님 공적 사항 보니까 전산 쪽 사전이 하나 있더라고요.

최기선 저기 보이는 '다국어 어휘 의미망'인데요. 중국어, 일본어, 한국어,

영어가 있고, 숙명여대 교수께서 만든 독일어도 있습니다. 이게 우리나라에서는 최초인데요. 영어 워드넷(WordNet)과 똑같이 출발했습니다. 결국은 영어 워드넷이 중심이 돼서 거기 개념 번호와 우리 개념 번호를 일치시켜 놨어요.

이경우 컴퓨터 사전을 개발했다는 표현을 썼던데요.

최기선 네, 컴퓨터용 사전 개발이죠. 그게 사람이 읽을 수 있는 형태가 아닙니다. 온갖 기호가 들어가 있어요.

이경우 컴퓨터가 보는 사전인가요. 이해가 안 가더라고요. 도대체 컴퓨터가 보는 사전이 뭐죠.

최기선 컴퓨터가 이해할 수 있는 사전입니다.

이경우 사람만 사전을 보는 게 아니군요.

최기선 그 사전을 펴 봐도 사람은 읽을 수 없죠.

이경우 그러니까 한국어를 이해할 수 있는 컴퓨터용 사전인가요.

최기선 네, 그렇죠.

이경우 공문서들을 보면 문장이 매끄럽지 못한 것들이 많은데요. 이런 것에 대한 컴퓨터의 지도도 가능하겠네요.

최기선 미국이나 영국에서는 언어라는 것이 공공성을 추구하는 면이 있는데 복합명사여서 어려우면 다 풀어 써야 하고, 문장도 쉽고 간결하게 잘 만들어야 하는 원칙들이 있습니다. 그 원칙들을 컴퓨터로 만들어서 쉽게 쓸 수 있게 하는 부분과 제약 언어와 관련된 부분에 대한 워크숍을 지난주에 주최했죠. 제가 국제표준화기구(ISO) 일도 하는데요, 여기서 '언어자원운영'도 제가 창립했습니다.

이경우 표준화라는 것이 무엇을 표준화하는 건지요.

최기선 표준에는 무엇을 만들기 위한 하나의 절차, 가이드라인에 대한 표준이 있고요. 또 명세에 대한 표준이 있습니다.

이경우 명세요.

최기선 네, 시맨틱 웹은 하나의 어떤 방식으로 하잖아요. 예를 들어 에이치티엠엘(HTML)을 어떻게 쓴다고 하는 방식이죠. 이것을 어떻게 쓰는지에 대한 표준이 있고, 글을 어떻게 쓰면 된다든지, 전문용어 같으면 전문용어 개발 원칙에 대한 표준이 있는 거죠. 그러니까 두 가지입니다. 직접적인 대상에 대한 표준이 있고, 그것을 만들기 위한 절차에 대한 표준이 있습니다. 그리고 용어나 심벌에 대한 표준도 있고요. 제가 하는 '언어자원운영'에서는 절차 표준이 많고요. 원칙이 있고 언어 분석을 했을 때 분석의 결과는 어떤 형태가 돼야 한다는 기준이 있는 거죠.

이경우 그곳에서 하는 일이 어떤 용어에 대한 표준보다 컴퓨터 언어의 표준이라고 봐야 하는군요.

최기선 그렇죠. 컴퓨터가 처리했을 때의 분석 결과 형태는 어떠해야 하고 분석하는 절차는 무엇 무엇이 있는 것을 추천한다고 돼 있는 것이지요.

이경우 여기 플러그가 있는데 이것의 규격을 모두 맞춰야 하는 것과 같은 건가요.

최기선 소켓으로 치면 우리나라는 2구로 돼 있고 일본 같으면 일자로 돼 있습니다. 이런 게 형태 자체에 대한 표준이고요, 플러그를 만들 때 절연체가 얼마만큼 들어가야 하는지 추천하는 것은 절차 표준이죠. 저는 언어에 대한 것을 하니까 해석을 하면 뒤에 동사라고 쓸 거냐, 어떻게 하고 동사라고 할 거냐 등 결과에 대한 명세를 하는 것이지요.

이경우 아까 유아르아이를 말씀하셨는데 유아르엘(URL: Uniform Resource Locator)은 많이 들어 봤어도 유아르아이는 생소합니다.

최기선 유아르엘은 어드레스, 그러니까 로케이터(locator)이고 유아르아이는 아이덴티파이어(Identifier)입니다. 최기선 하면 최기선에 대

한 유아르아이를 줘요. 또 다른 최기선도 있을 수 있는데, 그러면 '최기선 1'이다 하면 저고, '최기선 0'이다 하면 다른 사람이 되는 거죠. 웹에서 식별할 수 있는 주민등록번호 같은 거죠. 기계가 자동적으로 최기선에다 번호를 부여해 가는데 컴퓨터가 자동적으로 인덱스를 만들어 가는 거예요. 저에 대한 최기선도 있고 문장 속에서 어디선가 선생님이라는 말도 있고 했을 때 의미적인 색인을 만든다면 첫째 줄의 최기선과 25번째의 선생님은 똑같이 지식 베이스에서 '최기선 1'이라고 인덱스를 만들어 주게 되죠. 이렇게 되면 트위터나 신문에서 어떤 최기선을 써도 컴퓨터가 다 알아서 해주는 거죠.

이경우 누군가가 새로 태어나서 최기선이란 이름을 갖게 되면 어떻게 되죠.

최기선 새로 태어난 아기가 지금까지 나온 최기선과 다른 양상을 보이면, 새 개체로 인식해 새로운 엔티티(단위)를 넣어 주죠. 컴퓨터가 알아서.

이경우 쉽지 않습니다. 국어를 전공하지는 않으셨지만, 누구보다 국어학에 대한 지식도 대단하신 것 같습니다. 학교 다닐 때 국어 과목도 좋아하셨겠습니다.

최기선 고등학교 때 국어를 잘했습니다. 고문도 좋아했고요. 이 공부를 하면서 좀 하면 되겠더라고요.

이경우 전문용어와 관련해서도 많은 일을 하신 것으로 알고 있습니다. 전문용어에 대해 관심을 갖게 된 배경은 무엇인지요.

최기선 1990년 텍스트 분석을 하다가 텍스트 색인 만드는 방법을 독일에서 발표하게 됐어요. 그런데 일본 전문용어협회에서 발표를 보고 일본에 한번 오라고 하더라고요. 그래서 갔더니 한·중·일 전문용어가 링크되면 좋을 거 같다는 얘기를 하더군요. 처음 듣는 거였죠. 정말 중요하겠더라고요. 무슨 얘기를 하는지 서로 알아들으면

정말 좋겠다는 거였죠. 국내 배터리 가게에서도 일본 말로 '마후라'라고도 하고 '머플러'라고도 씁니다. 이렇게 용어 통일이 안 되던 상황이었는데 '21세기 세종계획'을 진행하다가 전문용어센터를 만들게 됐지요. 그리고 그걸 표준화하게 됐죠. 저는 자연어 처리를 하면서 용어가 중요하다는 생각을 하게 됐는데, 전문 분야에 들어가서 용어를 모르면 이해할 수가 없거든요. 이것 때문에 일단 시작을 하게 됐습니다. 기계 매뉴얼 같은 곳에 해설을 달아야 하는데 안 되는 거예요. 전문용어는 지금도 안 되는 게 많이 있어요. 사전이 없어서 그렇지요. 이게 첫째 이유고요, 둘째 이유는 학술용어와 산업용어, 표준용어가 서로 연결이 안 돼요. 예를 들어 '디프 러닝(Deep Learning)'이 나오니까 그냥 쓰잖아요. 디프 러닝을 심화학습이나 심층학습으로 바로 쓰지는 못하는 거죠. 그런데 디프 러닝이라고 했을 때 이 말을 우리나라 사람들이 과연 얼마나 이해할지 모르겠어요. 영어 쓰는 사람들은 너무 쉬운 거죠. 우리나라 사람들은 왜 심화학습이나 심층학습 같은 말을 사용하지 못할까요. 한국어로 돼 있지 않으면 용어가 우리에게 주는 효과는 없는 거나 마찬가지입니다.

이경우 일상생활에서도 전문용어가 미치는 영향이 상당할 텐데요.

최기선 상당히 많이 있습니다. 영어와 일본어에서 전문용어 통계를 보면 영어에서는 일상용어와 전문용어가 겹치는 비율이 약 70%나 됩니다. 사람들이 일상에서 쓰는 용어와 전문 분야에서 사용하는 용어가 상당수 같으니까 일반 국민의 수준이 높아지는 거지요. 일본도 이보다는 못하지만 40%는 똑같다고 합니다. 일본에서 낸 통계지요. 우리나라는 이런 게 체계화돼 있지 않은 거죠. 그런 생태계가 안 돼 있는 거예요. 디프 러닝이라고 하면 분명하게 알아듣는 사람이 얼마나 되겠어요.

이경우 전문용어가 더 쉬워지고 일상용어에 가깝게 된다면 국력이 커지고 사회 수준이 높아지는 거겠죠?

최기선 당연한 말씀이죠. 스스로 생각할 수 있는 힘이 커지는 거잖아요. 그러면 얼마나 파급 효과가 커지겠어요. 2002년 일본에서 노벨화학상을 받은 화학자는 잘 알려지지 않은 대학을 나왔고, 조그만 기업에서 실험 장비를 만들던 사람이잖아요. 일본에 가면 일본어로 다 교육하지 영어 교과서 안 쓰더라고요. 그러니까 그 친구는 자생적으로 자기 생각을 표현할 수 있는 거지요. 말이 이해되고 와 닿으니까 그런 일을 한 거죠.

이경우 전문용어들이 일본어로 돼 있으니까 일본어로 이해하고 자기 생각을 하게 돼서 상을 받은 거란 말씀이군요.

최기선 그렇죠. 노벨상도 영어 잘하면 되는 거 아니냐고 그러는데 한국 사람이 영어로 추론하는 것보다 한국어가 훨씬 더 낫죠.

이경우 언어 규범도 중요하지만, 전문용어 분야 또한 이에 못지않게 중요한 분야 같습니다.

최기선 우리가 한국어 자연어 처리를 할 때 전문 분야 특허 해설을 해야 하고, 교과서에서도 전문 분야 해설을 해야 하거든요. 그런데 물리학도 있고 화학도 있고 세계사도 있잖아요. 그럼 그걸 다 이해해야 되잖아요. 그런데 지금은 개념들에 대한 정리가 잘 안 돼 있기 때문에 힘듭니다.

이경우 전문용어 분야에서도 선생님께서 선구자적인 역할을 하신 거네요.

최기선 제가 처음 시작할 때는 하시는 분들이 별로 없었어요. 그래서 오스트리아 빈에 계시던 크리스티안 할렌스키라는 분을 찾아가서 배웠습니다. 이후에 교육부나 문화체육관광부에서도 이 분야의 중요성을 알고 여러 방면으로 시작했죠. 그렇지만 원칙이 없는 상태에서 일을 하게 됐죠. 각계에서 만들어 놓은 덩어리들이 있습니다.

이것들이 각급 학교 수준에 맞게 교과서에 들어가야 하는데 뒤죽박죽인 상태가 된다는 얘기를 많이 들었습니다. 결국 전문 분야가 정리돼야 컴퓨터가 텍스트를 다 이해할 수 있는 경지가 되는 거죠. 번역도 마찬가지고 검색도 마찬가지인 거고요, 지식정보도 마찬가지일 거고, 국민 수준 향상 이런 것도 그렇고요. 지금 당장 닥친 것은 학술적으로 막 등장하는 디프 러닝이나 산업화되는 특허에 쓰는 용어들도 있고요. 이게 정리가 안 됐기 때문에 한쪽에서는 파란색으로 보이고, 다른 쪽에서는 하얀색으로 보이게 됩니다. 그래서 혼동이 일어나게 되는 거죠.

이경우 신문이나 방송에 나올 정도면 일반 대중도 어느 정도 감을 잡을 수 있어야 하는데, 외국어가 대개 그대로 들어오는 거잖아요.

최기선 하나의 생태계를 만들고 이걸 다룰 수 있는 체계가 있어야 합니다. 학계에서 준비가 돼 있어야 하죠. 학계가 이걸 하기에는 너무 바쁘거든요.

이경우 1998년 전문용어 작업을 시작한 이후 많은 시간이 지났습니다. 그동안 여러 제안도 하셨을 거고요. 그런데 그 생태계가 아직 정리되지 않고 있는 거죠. 이것은 정부가 해야 하는 거잖아요. 국어원이 나서야 하나요.

최기선 과학기술 분야는 여러 연구재단도 있고 한국 과학기술한림원 같은 곳도 있습니다. 그런 곳에서도 일할 수 있는 부서가 커지고 정부도 지원을 하고, 저 같은 사람은 생태계를 만들고 그래서 그것이 작동되도록 해야 합니다.

이경우 이제 기대하시는 만큼의 인식 변화가 됐다고 할 수 있는지요.

최기선 여러 번 시행착오가 있었고요, 지금은 이러한 것에 대한 이해가 많이 되고 있습니다. 이러한 밑바탕에 대한 알맹이들이 제대로 정리되고 잘 쓰이도록 해야죠. 인공지능 같은 사업과 병행돼야 하는

데, 많이들 도와주시는 것 같습니다. 분위기는 조금 달라진 것 같아요.

이경우 아직 좀 형식적이지 않은가요.

최기선 이 분야가 융합적이잖아요. 문체부도 있고 미래창조과학부도 있습니다. 연결해서 해야 하는 일인 것 같습니다. 용어가 표준화된다면 공부하는 사람에게도 얼마나 좋겠어요.

이경우 결국 과학기술이 발전하려면 국어가 밑바탕이 돼야겠습니다.

최기선 뭔가 연결될 색인이 있어야 찾을 수 있는 거죠. 이걸 찾아 나서야 하는 수고가 필요한 상태입니다. 개념화되어 논리적으로 정리가 돼 있으면 기계가 이해하고 추론할 수 있는 능력이 커지기 때문에 웹 문화, 웹 과학 등 우리가 과학 전 분야에서 앞서게 되는 거죠.

이경우 선생님께서 말씀하신 것들이 이뤄지려면 과학계에 계신 분들이 더 적극적으로 말씀해 주셔야 할 것 같습니다. 전문용어 정리의 중요성에 대해서 말입니다.

최기선 앞에서 말씀드린 일본의 그 친구가 노벨화학상을 받았듯이 전문용어가 쉬워지고 표준화되면 할 수 있는 게 굉장히 많아집니다. 그런 사람들이 노벨상을 받으려면 용어가 더 낮은 수준까지 내려가야 합니다. 디프 러닝이라고 하지 말고 누구나 이해할 수 있는 수준까지 가야 하는 거죠. 말이 쉬워져야 하고 교육이 거기에 맞춰서 가야 합니다.

이경우 선생님께서는 남북한 전문용어 표준화 작업까지도 하셨다고 들었습니다. 잘 진행되고 있나요.

최기선 다른 방향에서 사전 만드는 것은 진행되는 것 같아요. 1995년도께 중국 연길시에서 '코리안언어학회'라는 것이 열렸죠. 거기서 전문용어에 대해서 하니까 북쪽에서도 그걸 하시는 분이 나오셨죠. 거기는 전투적인 용어를 많이 쓰더군요. '스택'이라고 있는데 한쪽

에서 다른 쪽으로 빠져나가는 거죠. 이것을 거기서는 탄창이라고 사용하더군요. 우리나라도 뭐 대책이 없는 거지만요.

이경우 일본이나 미국에서는 일찍이 중요성을 인식하고 있었네요.

최기선 그렇죠. 일본에서는 메이지유신 때 용어 통일을 한번 했고요, 1930년부터 학술용어는 표준화를 계속하는 거예요. 일본은 일을 벌일 때마다 용어 표준화를 먼저 하죠.

이경우 이외에 관심을 갖고 연구하시는 분야는 무엇인지요.

최기선 지금은 주로 한국어 기계 독해예요. 머신 리딩이라는 거죠. 디비피디아라는 것이 있어요. 위키피디아 테이블의 정보들을 디비(DB)화한 게 디비피디아예요. 예를 들어 위키피디아 한 페이지를 보면 여기에 버락 오바마가 있고 글이 있고 오른편에 테이블이 있어요. 오바마의 생일은 언제고 정당은 뭐고 이런 식으로 돼 있는데, 이를 떼어다가 디비화하는 것이 디비피디아예요. 한국어판은 우리가 하고 있어요. 이것을 지식 베이스라고 하죠. 1차적인 지식 베이스예요.

이경우 언어공학을 하시면서 가장 아쉬운 건 무엇인지요.

최기선 교수로서 논문을 발표하고 인정받는 것은 중요한 일입니다. 한데 한국어 데이터를 가지고 발표하면 잘 안 됩니다. 영어만큼 고도의 결과를 보이기가 어렵습니다. 왜냐하면 데이터가 워낙 적으니까요. 그렇다 보니까 학생들이 데이터로서 한국어를 안 쓰려고 합니다. 영어를 쓰려고 하죠. 한국어도 영어만큼 실험을 할 수 있는 데이터나 알고리즘 같은 것이 발전해야 합니다. 그래서 우리 학생들이 자신이 쓰는 말로 자신 있게 발표할 수 있는 세상이 됐으면 좋겠습니다. 이것이 교수로서 첫째로 해야 할 일이라고 생각합니다. 이것을 하기 위한 교과서나 기본적인 데이터나 실험 장치들은 제가 꼭 어느 정도까지는 만들어 놓고 싶습니다.

이경우　네, 오늘 좋은 말씀 잘 들었습니다. 고맙습니다.

최기선　감사합니다.