

솔트룩스, 국립국어원과 '2020 일상대화 말뭉치 구축' 사업 수주

[2020-05-19] 인공지능·빅데이터 선도기업 솔트룩스(대표이사 이경일)가 국립국어원과 사투리를 포함한 일상 대화를 빅데이터로 만드는 사업에 착수했다고 19일 밝혔다.



솔트룩스와 국립국어원은 '2020 일상대화 말뭉치 구축' 사업으로 일상 대화 속 언어를 인공지능에 활용 가능한 데이터로 만든다는 계획이다.

해당 사업은 2019년부터 진행한 '국가 말뭉치 구축사업'의 일환으로 ▲구어 ▲일상 대화, ▲메신저 대화 ▲웹 ▲문어 등으로 나눈 사업 분야 중 일상 대화에 속한다. '말뭉치'란 언어 연구 분야에서 언어 자료를 나타내는 용어로 연구 목적마다 구체적인 정의가 다르나 이번 사업에서는 약 500시간의 대화가 말뭉치 자료의 기준이다.

솔트룩스는 약 2,000명의 대화를 수집하고, 이를 고품질의 AI 데이터로 구축한다. 초기 단계에는 대화형 인공지능 분야 전문가들이 참여해 어떤 대화를 수집할 지 주제를 선정한다. '대화형 인공지능'이란 사람과 유사한 수준의 대화가 가능한 인공지능으로, 말뭉치 데이터가 주로 사용될 분야이다. 또한 대화를 수집할 때 방송에서 쓰이는 표준어보다는 실제 지역색이 그대로 묻은 대화를 수집하는 데 비중을 둔다.

수집한 음성 대화는 데이터로 가공하기 위해 텍스트 형식으로 옮긴다. 이 과정에서는 국내 최고의 속기 전문기업 '소리자바'와 협력한다. 텍스트 자료는 인공지능 활용 가능한 데이터 파일로 가공되어 국립국어원이 요구한 메타 정보가 함께 실어질 예정이다. 또한 솔트룩스에서 말뭉치 데이터를 활용할 수 있는 음성인식 기능을 탑재한 모델을 제공한다.

국립국어원은 "말뭉치 데이터를 민간에 공유해 언어 인공지능 등 관련 산업 기반을 마련하는데 기여할 수 있을 것"이라고 설명했다.

문의

- IPO/재무: 김형경 팀장(02-2193-1616) / fm@saltlux.com
- 회사소개: 서승역 팀장(02-2193-1603) / ccomkt@saltlux.com
- IR큐더스 이상호 책임(070-5068-6324)

말뭉치 데이터는 인공지능을 활용한 서비스 개발 과정에서 학습 데이터로 활용될 가능성이 있다. 학습 데이터란 인공지능이 목적대로 작동하기 위해 학습하는 데이터로, 기계어 형식으로 만들어져 있다. 따라서, 인공지능에게 사람의 말을 학습시키려면 기계어 형식의 학습 데이터를 만들어야 한다. 국립국어원에서는 이를 위한 기초 데이터를 만들어 인공지능 연구·개발에 힘쓰고 있다.

솔트룩스 이경일 대표는 "현재 세계 대화형 AI 시장보다 국내 시장은 음성 및 전사 말뭉치가 부족하고, 기술 개발이나 서비스화가 더디게 발전하고 있다"라며 "이번 사업을 통해 대화형 AI 산업 활용을 위한 기반이 마련돼 다양한 AI 서비스와 데이터 생태계가 확보될 것으로 기대한다"고 밝혔다.

한편, 솔트룩스는 대한민국의 대표 AI 기업으로, 국내 인공지능업체 중 가장 많은 121건의 기술 특허(PCT 포함)를 보유하고 있으며, 아시아 최대 규모인 150억 건 규모의 지식베이스를 구축해 인공지능 및 빅데이터 분야 기술 개발을 선도하고 있다.

회사는 지난 6일 코스닥 상장을 위한 예비심사를 통과했다고 밝힌 바 있다.

문의

- IPO/재무: 김형경 팀장(02-2193-1616) / fm@saltlux.com
- 회사소개: 서승억 팀장(02-2193-1603) / ccomkt@saltlux.com
- IR큐더스 이상호 책임(070-5068-6324)