NLP without NL Knowledge?

2017320114 컴퓨터학과 최재원

이번 수업은 임희석 교수님의 Natural Language Processing(자연어처리)와 관련해서, 자연어처리란 무엇인가 그리고 최근 딥러닝 기반에 NLP를 많이 하고 있고, 현재 NLP 전공자가 아닌 사람들도 NLP를 많이 하는 상황이 생겼는데 그러면 과연 NLP에 대한 공부는 필요가 없는 것인가, 더 나아가 자연어처리 연구의 중요성에 대해서 교수님의 세미나를 들어볼 수 있는 시간을 가졌습니다. 현재 사람이 컴퓨터보다 잘하는 것을 컴퓨터로 하게 하려는 것이 인공지능이라고 할 수 있다. 여러가지 인공지능 기술 안에서 말을 이해하고 말을 생산하는 것을 우리는 Natural Language Processing이라고 하고, 무엇을 경험을 통해서 학습을 자연스럽게 하는 것을 발전하는 것을 우리는 Machine Learning이라고 한다.

자연어처리란 컴퓨터를 이용해서 사람의 언어를 understanding하거나 사람의 언어를 generating하는 것을 기반한 연구를 자연어처리라고 한다. 사람들이 사용하는 일상적인 언어인 자연어를 컴퓨터에 입력 되었을 때 그 의미가 무엇인지 이해하는 것이 Natural Language Understanding이라고 한다. 그리고 컴퓨터가 의미가 주어졌을 때, 그 의미를 언어로 만들어 내는 것을 Natural Language Generation이라고 한다. 따라서 Natural Language Processing은 크게 이 두 가지를 하는 연구라고 생각하면 된다. 우리가 흔히 쓰는 Google Translation System(Google Translator)가 자연어처리 시스템의 대표적인 예시이다. 요즘 딥러닝 기법을 적용한 machine translation system이 만들어지면서, 기계 번역 분야의 아주 혁신적인 성능 개선이 있었다. 즉, 웬만한 언어들을 Google translator에 넣어서 번역을 하면 무슨 말을 하는 지는 대충은 얘기할 수 있을 정도로 만들어져 있다. 또 다른 응용 시스템은 챗봇이다. 이것은 대화 시스템으로 자연어처리의 대표적인 응용 영역이다. 이 챗봇은 다양한 분야로 진출할 수 있다. 예를 들어 콜센터에서 상담원 대신 챗봇이 전화를 받아서 사람들의 질문에 응대를 하면서 24시간 내내 쉬지 않고 고객들의 질문을 응대할 수 있는 장점을 가진다. 이렇듯 글(text)를 이용해서 하는 모든 영역은 자연어처리 영역이라고 생각하면 된다.

2017320114 최재원

NLP는 요즘은 pipeline system보다는 딥러닝 기반의 end to end 기법을 많이 사용하는데, 그렇다고 해서 pipeline system의 내용이 불필요한 것이 아니다. 자연어처리는 pipeline system의 단계별로 각각 연구 주제들을 가지고 있는데, 이것에 대한 접근법은 symbolic approach와 empirical approach가 있다. Symbolic approach는 knowledge-based, rule-based 등이라고도 불린다. 이 접근법은 입력이 들어오면 그 입력을 가지고 원하는 결과를 만들어 내기 위해서 사용할 수 있는 규칙들을 미리 만들어 놓는다. 문장을 이해하기 위해서 어떤 단어가 어떤 문맥에서 나오면 어떻게 해석할지, 또 어떤 문장이 어떤 문맥에서 나오면 어떻게 해석할지, 이런 류의 규칙들을 전문가들이 만들어 낸다. 그리고 입력이 들어오면 현재 이 입력에 적용할 수 있는 규칙이 무엇인지 답변을 만들어 내는 것이 symbolic approach이다. 이 접근법은 구현은 간단하지만, 단점은 우리가 어떤 문제 해결을 하는데 있어서 필요로 하는 모든 규칙을 미리 다 만드는 것이 불가능하다. 따라서 규칙으로 미리 만들어 놓지 못했던 입력이 들어오면 맞지 않을 수 있다. 따라서 사업성이 있는 아이템 보다는 토이 시스템을 만드는 것에 적합하다고 한다. Empirical approach는 statistical, probabilistic approach, 경험주의적 approach라고 할 수 있다. 마치 사람이 경험을 통해서 무엇을 학습하는 것처럼 기계도 학습한다고 보면 된다. 이 기법이 symbolic approach보다 많이 사용되고 있고, 딥러닝 기법도 이 방법에 속한다고 보면 된다. 하지만 이 approach의 단점은 data driven approach이기 때문에 많은 학습 데이터들을 필요로 한다. 또한 결과에 대한 해석 능력이 많이 떨어진다. 장점은 사람이 많은 노동력을 들여서 개발할 필요가 없다. 또한 확장성(coverage)가 높다고 할 수 있다.

이렇듯 현재 자연어처리의 중요성과 어떠한 기법들이 있는지 배울 수 있는 시간이 되었다. 이 분야가 조금 더 가까이 와 닿았던 점은 바로 현재에서 우리가 사용하고 있는 자연어들을 기계가 처리한다는 점에 있었다. 평소에도 인간의 말을 하는 AI들, 번역을 해주는 translator 등 여러가지 기술의 발전을 보면서 호기심을 가졌었는데, 교수님의 세미나를 통해서 조금이나마 원리를 배울 수 있었고, 현재 진행되는 연구가 무엇이 있는지 알 수 있었다. 평소에도 AI에 관심이 많았기 때문에, 최근 콜로퀴움 세미나에서 AI의 여러 기술 분야들에 대해서 강연을 해서 매우 열심히 듣고 있으며, AI도 세분화되어서 여러 분야가 있다는 것을 느낄 수 있었다. 개인적으로 기회가 된다면 자연어처리 수업을 다음 학기에 수강을 하고 더 배우고 싶다.