아이리포 지식창고

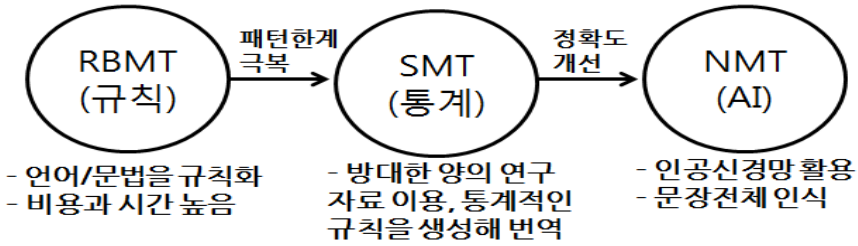
인공지능2 - 02 인공지능과 NMT

Neural Machine Translation

최창환 정보관리기술사
(buksamfight@naver.com)

인공지능 기반 통문장 기계번역!

Concept	**(정의)** 기계 번역 방식의 한 종류로 인공지능(AI)이 데이터 학습을 통해 문장 단위로 언어를 번역하는 기술
KeyWord	RBMT, SMT, NMT, RNN, Word Embedding, LSTM



네이버 파파고(Papago)를 아시나요?

최근 재미난 사건이 있었습니다. 바로 '인간 Vs 인공지능(AI)!' 번역 대결이 지난 2 월에 있었던 것이죠. 인간대표인 인간번역사팀과 인공지능 대표인 구글 Translation, 네이버의 파파고, 시스트란의 PNMT 가 대결을 펼쳤는데요. 이전 알파고처럼 큰 반향을 일으킨 건 아니지만 저는 이 소식이 상당히 궁금했고 결과를 숨죽여 지켜봤습니다.

이유는 알파고 때처럼 또 한번 전 세계를 뒤흔드는 결과가 나오는게 아닐까 하는 기대감 혹은 우려 였고, 그리고 국내의 인공지능 기술 수준도 간접적으로나마 볼 수 있다는 생각 때문이 였습니다.

결과는??

많은 전문가들의 예측대로 인간의 완승으로 싱겁게 끝이 났습니다. 인간 번역사 팀은 한글을 영어로 번역한 문제에서 30 점 만점에 24 점을 받았고, 영어를 한글로 번역한 문장은 25 점을 받았습니다. 반면

Figure 1 구글번역과 네이버파파고 특징

인공지능이 한글을 영어로 번역한 문장은 13 점, 7 점, 8 점을 받았고, 영어를 한글로 번역한 인공지능 점수는 각각 15 점과 8 점, 9 점을 받아 인간과의 격차를 확인 하게 되었습니다. 아직은 문맥의 자연스러움에서 인간통역사가 한 단계 높은 번역실력을 보여 주었으며, 인공지능의 학습 데이터 부족과 번역 완성도를 고려 해야 한다는 점에서 학습 시간이 좀 더 필요 한 거 같습니다. 그래도 한가지 놀라웠던 것은 이번 대결에 참여한 '인공지능 녀석?' 들이었는데, 향후 2020 년 80 조라는 큰 시장을 기대하고 있는 통번역 시장에서 국내의 순수 기술로 개발 된 **네이버의 파파고, 시스트란의 PNMT 가 세계적인 구글 Translation 과 어깨를 나란히** 하는 것을 보고 '와~ 정말 우리도??' 라는 기대감에 기분이 좋았던 것은 사실입니다.

잠시 기계번역에 대한 역사를 되짚어 보겠습니다.

기계 번역의 태동은 1950 년대 미국과 소련의 냉전시대 대립에서부터 시작했습니다. 당시 미국의 과학자들은 소련의 언어를 컴퓨터로 번역하기 위해 골머리를 앓아야 했습니다. 러시아어와 영어의 법칙을 풀어 규칙화 하면 컴퓨터가 이를 기반으로 번역할 수 있을 것이라 생각했습니다. 이때 미국이 활용한 기술이 **'규칙기반 기계 번역 (RBMT, Rule-Based Machine Translation)'**입니다. 규칙기반 기계 번역 기술은 **언어학자가 언어의 문법을 하나하나 전부 규칙화** 했기 때문에 번역 정확도가 높았으며 품질의 일관성과 예측가능성이 높았다는 것이 장점입니다. 그러나 높은 언어 지식을 지닌 언어학자가 언어의 문법을 일일히 규칙화해서 설계해야 하는 점은 장점임과 동시에 단점이었습니다. 문법을 규칙화하는 작업에 들어가는 비용과 시간이 상당히 크기 때문이라고 하는데요. 또한, 규칙화 되어 있는 패턴을 벗어나는 경우엔 번역을 처리하기 힘들어 지는 것이죠. 이후 1988 년 IBM 이 **'통계기반 기계 번역(SMT, Statistical Machine Translation)'** 기술을 기계 번역에 도입하였습니다. 통계기반 기계 번역은 **방대한 양의 연구 자료를 이용해 통계적인 규칙을 생성해 번역하는 방법**입니다. 그렇기 때문에 언어학자가 번역 대상 언어의 문법을 규칙화 할 필요가 없으며 규칙기반 기계 번역을 개발하는 데 들어가는 시간과 비용을 상당히 줄일 수 있었습니다. 또한, 이미 번역된 결과의 통계를 개발에 활용하기 때문에 개발 시간을 단축할 수 있고, 특정 언어에 국한되지 않는 시스템을 개발할 수 있습니다. 통계기반의 기계 번역은 데이터가 많이 쌓일수록 번역의 품질이 높아지는 것도 장점이라고 할 수 있습니다. 반대로 단점은 대량의 데이터가 쌓이기 전까지는 번역의 품질이 떨어지는 것이며 양질의 번역 결과를 얻기 위해 방대한 양의 데이터를 저장할 저장 장치가 필요하다는 것입니다. 게다가 문법을 반영하지 못하기 때문에 문법 구조가 확연히 다른 언어 간의 번역을 할 경우에는 정확도가 다소 떨어진다고 알려져 있습니다.

가장 최근에 개발되어 인간의 번역 수준까지 정확도를 높인 기술이 무엇 인가하면 여러분들도 잘 아시는 **'인공신경망 기계 번역 (NMT, Neural Machine Translation)'**입니다. NMT 는 인간의 뇌를 모방한 인공신경망을 활용하여 규칙기반 번역 기술과 통계기반 번역 기술에서 비약적인 발전을 한 형태인데요. 기존의 기술이 단어나 구문에 초점을 맞추었던 것과는 다르게 인공신경망 기계 번역 기술은 **문장 전체나 문서 전체를 통째로 번역**합니다. 다시 말해, 인간이 번역하는 방식대로 문맥을 분석하여 인간이 이해하는 의미에 가장 가깝게 번역하는 것입니다. 지금도 인간의 번역 능력에 70%에 가까운 정확도를 보이고 있는 NMT 는 주어진 언어 데이터를 통해 딥 러닝을 수행하기 때문에 처리한 언어 데이터의 양이 누적될수록 번역의 질이 나아집니다.

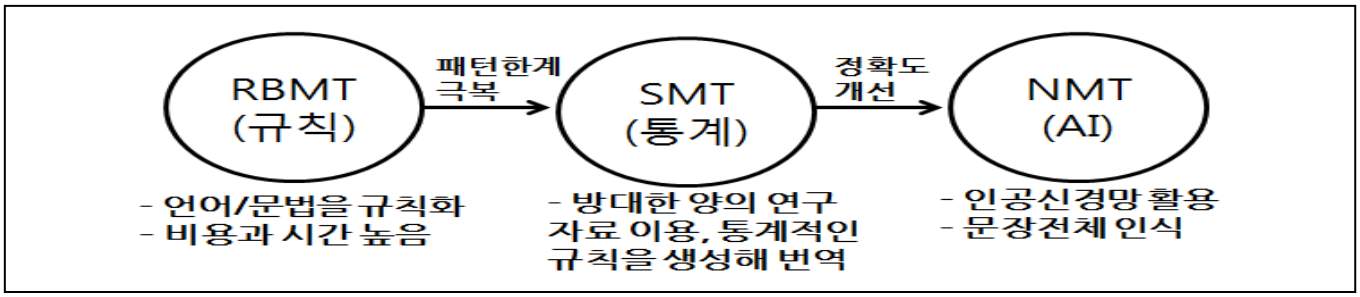


Figure 2 기계번역의 방식의 변화

이처럼 머신러닝 알고리즘인 NMT 의 성장으로 통번역 시장에서도 인공지능의 성장세가 뚜렷하게 나타내고 또한 성과도 보여주고 있습니다. 그렇다면 NMT 는 어떤 알고리즘으로 구성되어 있는지 한번 알아 보겠습니다.

'인공신경망 기계 번역 (NMT, Neural Machine Translation)'

'**기계 번역 방식의 한 종류로 인공지능(AI)이 데이터 학습을 통해 문장 단위로 언어를 번역하는 기술**. 기존 기계 번역에 주로 사용된 통계 기반 기계 번역(SMT, Statistical Machine Translation) 보다 발전된 형태다. 사용자가 번역 엔진에 **문장을 입력하면 맥락을 파악한 후** 이를 단어, 구문, 어순 등의 정보가 담겨 있는 벡터(좌표 값)로 전환하여 번역을 한다. SMT 와 달리 문맥을 파악할 수 있어 보다 자연스럽고 정확한 번역이 가능하다. **개발자가 신경망 구조를 결정해 주면 머신러닝, 딥러닝등을 통해 AI 가 스스로 학습하여 번역을 수행하므로 입력한 데이터가 많을수록 정교한 번역이 가능**하다. ' 라고 네이버에서 검색 했습니다.

가장 큰 특징은 문장 전체를 파악한 뒤 단어와 순서, 의미, 문맥 차이 등을 스스로 반영해 번역 하는 것입니다. 이러한 알고리즘은 어떻게 수행 하게 되는 것일까요? 참 궁금합니다. 머신러닝에서 영상인식에 탁월한 성능을 내는 **CNN(Convolutional Neural Network)** 알고리즘으로 먼저 생각해보면, 결론적으로는 입력 값에 대한 feature 을 통해 학습 된 분류기에 연관성에 대한 확률을 결과 값으로 도출 해 내는데, NMT 와도 전체 적인 알고리즘 면에서는 유사하다고 볼 수 있습니다.

NMT 는 문장을 통째로 입력 받아, 단어와 표현 값으로 변환하여 이 단어 표현들을 이어가며 최적의 가중치(weight parameter)을 찾아 행렬 곱으로 이어가며 벡터를 구해가는 방식입니다. 약간 수학 스러운 단어들이 나오기 시작하니까 갑자기 읽기가 싫어 지는 독자분들도 있을 거라 생각됩니다만.. 사례로 한번 살펴 보겠습니다.

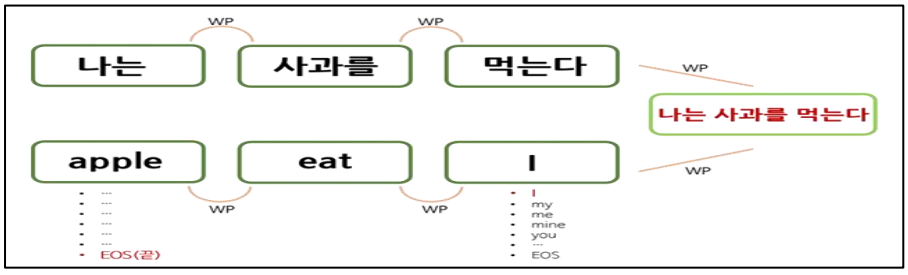


Figure 3 열쇳말 딥러닝과 기계번역 – NMT

[Figure 3] 에서 보면 각 단어를 분리 했고 각 단어를 이어가며 가중치 (WP)를 통해 **'나는 사과를 먹는다'** = **'I eat apple'** 은 같을 확률이 가장 높다 라고 결과값을 리턴 했습니다. 이 과정에서 **2 가지 핵심기술**이 사용이 되는되요.

바로, 워드임베딩(Word embedding) 과 **순환 인공 신경망(RNN, Recurrent Neural Network)** 입니다.

첫 번째는 하나의 단어를 인공 신경망을 이용해서 벡터 공간상에 나타낼 수 있는 변환 된 값으로 만들어 내는 단어표현, 즉 **워드임베딩(Word embedding)** 입니다. 한 단어를 입력하면 인공 신경망을 통해 단어와 관계 있는 관련 단어들이 워드공간(Word Space) 상에 매핑이 되고, 이 단어 표현의 벡터 값이 되는 것입니다. 위의 사례로 예를 들면 '사과를' 이라는 단어를 워드공간에 올리고 그 근처에 '먹는다', '먹었다' 등의 '먹다' 라는 단어와 관계가 있는 단어들과 유사한 공간에 두면서 서로 관계를 맺으며 매핑을 하게 되는 것입니다.

두 번째는 순환 인공 신경망(RNN, Recurrent Neural Network) 입니다. 이 머신러닝 알고리즘은 문장구조에 따라 단어를 반복적으로 트리 구조로 매핑하여 압축하고 복호화하는 형식으로 문장을 번역 하게 되는데, 사상은 이렇습니다. 바로 memory 입니다. 가장 최근 값을 미리 저장 하고 있다가 입력 된 값에 대에 반복 가중치를 부여 하겠다라는 의미로 보시면 되겠습니다. 아래 그림을 보시면,

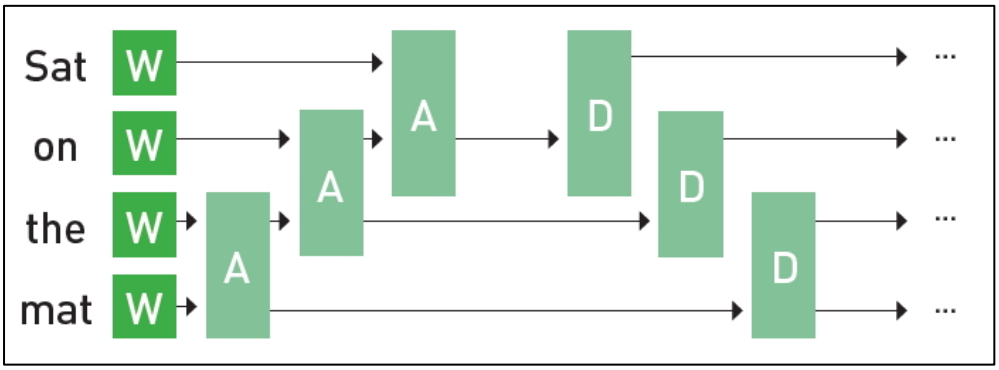


Figure 4 SPRi 기계번역 기술동향

W 는 워드임베딩 된 값이 각 차원에서 행렬의 곱에 따라 의미가 부연 된 값으로 변환이 됩니다. RNN 은 이 값을 시계열 데이터에서 패턴을 찾아내 압축(Encoder), 복호화 (Decoder) 하면서 출력과 얼마나 관련이 있는지를 검증하게 됩니다. 즉, A → D 가 얼마나 관계가 있는지를 가중치로 표현하게 되고 가장 높은 가중치를 출력으로 확정하게 됩니다. 물론 이 과정에서 수 많은 연산을 수행 하게 되는데 이때 활용하는 **알고리즘은 LSTM (Long Short Term Memory)** 이 가장 대표적으로 활용이 됩니다. 앞서 'RNN 은 가장 최근 값을 미리 저장한다' 라고 말씀을 드렸는데, 이때 활용 되는 녀석입니다. 즉, LSTM 의 이전 결과들을 저장하고 참조하여 기존의 LSTM 구조를 바꾸지 않고 신경망이 참조할 수 있는 정보를 더 늘려주는 효과를 가져올 수 있습니다. 정보량이 늘어 남에 따라 연산과정은 더욱 늘어나게 되지만 이 또한 GPU 를 각 Layer 별로 할당해 속도적인 측면을 개선해서 동시번역이라는 이상에 더욱 한발 더 다가가게 되었습니다. 구글(Google)의 GNMT 가 이를 이용하는 대표적인 사례이고 네이버의 파파고 역시 유사한 형태로 구성 되어 있습니다.

[표-1] NMT 기술요소

구분	기술요소	설명
Big Picture	**워드 임베딩**	- 한 단어를 입력하면 인공 신경망을 통해 단어와 관계 있는 관련 단어들이 워드 공간(Word Space) 상에 매핑이 되고, 이 단어 표현의 벡터 값이 곧 워드 임베딩
	순환 인공 신경망 (RNN, Recurrent Neural Network)	- 과거의 출력이 다시 입력이 되는 구조를 소위 피드백 구조 - 시계열 데이터에서 패턴을 찾아내는데 최적화된 방법론으로써 음성인식과 자연어 처리에 활용
Sub	Encoder	- 입력 된 문장을 기계가 이해하는 언어로 변환 하는 과정

Picture	LSTM	- RNN의 한 유형으로, 피드백 데이터를 기억하는 구조로 사용 (결과 저장, 참조)
	Decoder	- 기계적 언어를 변환하고자 하는 언어로 해독 하는 과정
	Soft Max	- encoder에 저장된 LSTM의 output들은 각각 softmax activation를 통해 관련도에 따라 확률 값이 계산

최근 기계번역 2020 년 약 80 조라는 거대한 시장을 이루게 되었습니다.

그리고 챗봇 등 다양한 서비스에도 접목을 시도 하고 있고, 일본에서는 로봇에 NMT 를 적용해서 통역 서비스를 준비하고 있다고 합니다. 국내에서도 앞서 말씀 드린 네이버의 파파고는 실시간 번역 서비스를 PC 환경과 모바일 환경에서 제공 하고 있고, SYSTRAN 의 PNMT 는 OCR(광학문자인식) 특허 출원을 통해 글로벌 시장 진출을 앞장서고 있습니다. 물론 아직 나가야 할 길은 많이 남아 있습니다. 정확한 문장 해석과 문맥 간의 자연스러움은 좀더 보완을 해야 할 것이고, 인간의 감성과 감정을 인공지능이 이해 하기에는 시간이 더 필요 할 것 입니다. 그렇지만 국내 업체들의 적극적 투자와 업체 간의 협력, 통번역을 위한 C-P-N-D 관점의 통합 전략 확대, 정부의 장기적인 지원과 산학연 협력 관계 구축 및 인공지능 확산을 위한 산·학·연 협력관계를 기반으로 기술개발에 집중한다면, 2020 년에는 국내 기업이 기계번역 시장 피라미드에 최상위에 올라 있지 않을까 조심스럽게 희망해 봅니다.

너무 극단적인 생각일지도 모르지만, 멀지 않은 시점에(전문가들은 향후 5 년 이내 예상)
우리는 외국어를 더 이상 공부 하지 않아도 되는 세상을 만날지도 모르겠습니다.

"끝"