기계번역기의 A to Z, 그리고 카카오의 기계번역

AI in Kakao

인공지능(AI) 기술이 우리 눈 앞에 구현되고 있는 대표적 형태 중 하나가 기계번역입니다. 이번 호에서는 기계번역이 발전되어 온 역사, 그리고 현재 기계번역 기술의 급성장을 가져온 신경망 번역 모델이 진화해 온 양상에 대해 설명한 글을 담았습니다. 카카오 번역기가 경쟁력을 높이기 위한 목적에서 양질의 대규모 학습 데이터 확보를 위해 사용한 기술에 대한 설명글도 준비했습니다.

AI in Kakao

기계번역기의 역사와 발전

신경망 기반의 기계번역(neural machine translation, NMT)은 메이저급 기계번역 서비스에 속속 도입되고 있다. 딥러닝과 방대한 병렬 말뭉치를 핵심 기술로 삼고 있는 기계번역기는 번역 전공자들의 위기감까지 유발할 정도의 높은 성능을 자랑한다. 이러한 시스템의 등장은 지난 70여년 간의 '번역'이라는 작업(task)에 대한 깊은 이해와 고민, 여러 차례의 시행 착오가 없었다면 불가능했을 것이다. 기계번역기의 발전 연대기는 기술의 발전 양상에 따라 (1) 1960년대 중반, (2) 1960년대 중반에서 1990년대 중반, (3) 1990년대 중반 이후로 구분될 수 있다.

글 | 김미훈 may.jin@kakaocorp.com

현재 14개월된 아기 엄마로서 인간지능의 발달과정과 인공지능간 차이의 분명함을 매일 경험하면서 인간을 모형으로 하고 있는 인공지능에게 '진정한 지능'을 만들려면 어떻게 해야 할까 항상 고민하고 있습니다. 밖에서 보여지는 인공지능 신화와 달리 아직은 할 일이 너무 많다고 생각하는 현업 종사자로서 매일매일 최선을 다할 뿐입니다.

기계번역기의 발전 연대기

기계번역이라는 용어는 1949년 워렌 위버(Warren Weaver)의 번역(translation)[*1]에 처음 언급 되었다. 이 논문은 당시 큰 주목을 받았고, 기계번역 연구의 촉매 역할을 한 것으로 평가된다. 1951년 MIT에서 기계번역 연구가 본격적으로 시작됐다. 이후, 일본, 러시아, 프랑스 등의 국가에서도 기계번역 관련 연구가 활발히 진행됐다. 다음의 유튜브 영상[*2]을 보면 1950년대 기계번역에 대한 기대 정도를 엿볼 수 있다.

[그림 1] 1950년대 기계번역

5년이면 기계번역을 정복할 수 있을 것이라는 기대와 달리, 1960년대 중반에 와서도 기계번역 성능은 기대에 미치지 못했다. 기대는 실망으로 바뀌었고, 기계번역에 대한 회의론이 일기 시작했다. 기계번역 연구도 이전처럼 활발하지 못했다. 관련 연구는 여러 대학 연구실에서만 주로 이루어졌다. 1970년대와 1980년대를 거치면서 여러 방법론들이 제시되었지만 일부 분야(domain)에서만 그 성능을 인정받는 수준이었다. 하지만 이 시대의 기계번역 연구는 관련 자연어처리 기초 연구 분야들(형태소 분석, 구문분석, 언어생성 등)의 발전에 많은 영향을 주었다.

[그림 2] 1970년대 기계번역

1990년 전후로 통계적 방법을 기계번역에 접근한 통계 기반 기계번역(statistical machine translation, SMT)은 혁신적인 변화를 일으켰다. 원문과 번역문이 함께 있는 병렬 말뭉치에 통계적으로 접근하는 방법론이다. 다수의 기업들이 이러한 방법으로 기계번역 개발에 적극 뛰어들기 시작했다.

이후 딥러닝 기반 방법론들이 이미지 처리, 음성인식, 자연어 처리 기반 기술로 사용되면서 기계번역도 좋은 성능으로 주목받게 되었다. 번역 관련 서비스들이 폭발적으로 늘어남에 따라 구글(Google), 마이크로소프트(Microsoft), 페이스북(Facebook) 뿐만 아니라 중국의 바이두(Baidu), 러시아의 얀덱스(Yandex) 등 현재 각 나라의 주요 포털 업체들의 경우 모두 자체 플랫폼에서 번역 서비스를 제공하고 있다. 이외에도 시스트란(SYSTRAN)과 같은 번역 서비스 회사들까지 활발하게 사업을 펼치고 있는 중이다. 그래서 지금은 가히 기계번역기의 춘추전국시대라고 해도 과언이 아니다.

아울러 해외 여행의 급증, '직구' 등을 통한 해외 쇼핑 경험 확대, 비즈니스 교류 증가 등으로 통번역 수요가 높아지고 있다. 모바일 플랫폼 확장으로 인해 높아진 번역 서비스에 대한 접근성도 기계번역에 대한 필요를 키우고 있다. 이와 같은 번역 서비스에 대한 수요의 지속적 증가는 기계번역기를 둘러싼 치열한 경쟁의 형성에 한 몫을 하고 있다. 이같은 환경 속에서 주요 IT 업체들은 자사의 기계번역 기술을 바탕으로 사용자에게 글의 맥락(context)에 최적화된 번역 서비스를 제공하려고 노력하고 있다. 아울러 적용 폭 역시 확대 중에 있다. 시장조사 기관인 그랜드 뷰 리서치(Grand View Research)에 의하면, 기계번역 시장은 2022년 9억 8330만 달러 규모(1조 747억 원[*3])로 성장이 예상된다. 특히 방대한 양의 콘텐츠를 정확하고 빠르게 번역할 필요가 있는 전자, 자동차, 의료, 밀리터리 업계 등에서 기계번역의 높은 미래 가치가 점쳐지고 있다[*4].

구글, MS, 바이두의 기계번역

현재 전 세계적으로 가장 많이 이용되고 있는 기계번역 서비스는 구글 번역이다. 구글 번역은 103개 언어의 번역을 지원하고 있는데, 이는 세계 최대 규모다. 구글 번역을 이용하는 사람은 전 세계 5억 명 이상이며, 매일 1400억 개 이상의 단어 번역이 이루어진다.

구글은 2006년 통계 기반 기계번역 서비스를 출시한 후, 다음 해 모든 번역 엔진을 SMT로 전환하였고, 2016년 10월 GNMT(Google's neural machine translation)를 발표하며 이를 서비스에 적용하기 시작했다. 또한 구글은 증강현실 애플리케이션 서비스 업체인 '퀘스트 비주얼(Quest Visual)'을 인수해 이미지 번역을 시작했다. 최근에는 40개 언어를 자동으로 번역해주는 구글 어시스턴트(Google Assistant)가 내장된 무선 헤드셋 '픽셀 버즈(Pixel Buds)'를 출시하며 하드웨어 분야로의 진출을 본격적으로 알리고 있다.

MS는 2016년 11월에 기계번역 시장 경쟁에 뛰어들었다.

기술문서 번역의 강점 및 엔터프라이즈(enterprise)[*5] 서비스 경험을 기반으로 한, 발표자의 설명을 실시간으로 번역하는 프레젠테이션 번역, 스카이프(Skype)를 통한 실시간 통번역 서비스 등을 제공하며 기업 시장을 공략하고 있다.

중국의 구글이라 불리는 바이두는 정부의 적극적인 지지에 힘입어 최근 인공지능(artificial intelligence, AI) 분야에서 가파른 성장세를 보이고 있다. 지난 2년간 바이두가 AI 분야에 투자한 금액은 200억 위안(약 3조 6464억원)[*6]에 육박한다. 바이두는 강력한 데이터베이스를 기반으로 현재 27개 언어에 대한 번역 서비스를 제공하고 있으며 일부 언어에 NMT를 적용하고 있다.

국내의 기계번역 시장 역시 IT 기업들이 개발을 주도하고 있다. 카카오는 지난 10월 통합 인공지능 플랫폼, 카카오i(아이)의 번역 엔진을 적용한 기계번역 서비스 '번역 베타(beta)' 서비스를 선보였다. 한국어, 영어 이외의 추가적인 언어 서비스를 위해 준비 중이다. 네이버는 '파파고(papago)'라는 서비스를 운영 중이다. 2016년 8월부터 운영 중인 파파고에는 NMT 기술이 적용됐다.

높은 관심과는 별개로, 기계번역 역시 딥러닝 연구의 빙하기와 마찬가지로 인기가 없던 암흑기가 있었다. 인기 없는 연구임에도 불구하고 이 문제를 해결하고자 시간과 심혈을 기울였던 모든 이와 연구를 지원해준 관련 정부 부처 연구비 집행 담당자분들의 안목이 있었기에 오늘 이러한 번역기를 만들 수 있다고 생각되기에 그 모두에게 감사드린다.

번역 기술의 개괄적 구조

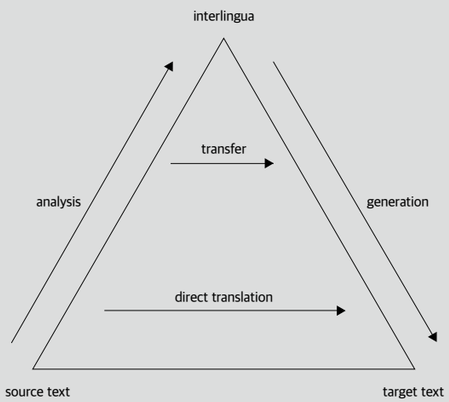
번역 기술을 [그림 3]과 같은 형태로 표현해볼 수 있다. 번역하려는 언어(source language)와 번역언어(target language)를 삼각형 아래 두 꼭지로 표현할때, 번역 과정을 크게 3가지 경로에 따라 direct, interlingual 그리고 transfer방식으로 표현할 수 있다.

첫번째 경로가 보여주는 것은 소스 언어(source language)에서 타겟 언어(target language)로 direct-translation이라고도 하는데, word-to-word 번역 방식이다. 초창기에는 이 방식이 한국어와 일본어처럼 서로 유사한 언어를 번역하는 기계에 사용되기도 했다. 이 방법론의 장점은 간단함에 있지만, 단점은 토폴로지(topology)가 다른 언어쌍의 번역을 진행하기에는 부족하다는 것이다.

Source-target 언어 쌍의 차이를 극복하기 위한 방법으로 학자들은 source 문장들의 의미를 추상화하여 원래 언어와 독립된 매체 또는 구조로 표현하고, 이 추상화된 의미를 다시 target 언어로 생성하는 방식으로 접근했다. Source 언어 문장을 interlingua로 표현하고 다음으로 interlingua에서 target언어 문장을 생성 할 경우, 우리는 이를 interlingual 방식으로 부른다.

반면 transfer는 direct-translation과 interlingual-translation의 중간 정도 abstraction으로 생각할 수 있는데, 예로 syntactic transfer에서는 source 단 문장들을 (의존)구문분석[*7]을 통하여 구문트리로 표현하고, 구문트리는 변환과정을 거쳐서 target 번역문을 만드는 과정이다.

[그림 3] direct-, interlingual- and transfer-translation



1990년대 부터 말뭉치를 이용한 통계적 접근 방식이 자연어처리 등 여러 task에 적용되면서, 병렬 말뭉치를 이용한 통계 기반 기계번역 방법론들이 통용되기 시작했다. 통계 기반 기계번역은 translation model(이하, TM)과 language model(이하, LM)로 이루어져 있다. TM에서는 source와 target 문장 사이에 각 단어나 구가 어떻게 번역으로 매치되는지를, LM에서는 번역문의 각 단어들이 얼마나 문장다운 문장을 만드는지 수치화하여 보여준다. 이 두개 요소가 함께 번역하려는 문장에 딱 알맞는 번역문을 만드는 것을 제어한다.

NMT는 SMT의 범주로 볼 수 있는데, 신경망을 기반으로 TM이나 LM의 수치들을 얻는다는 점이 SMT와의 차이점이다. 최초 NMT[*8] 모델의 중심은 LM이었고 그 이후 제안된 모델[*9]에는 attention이라는 mechanism으로 source와 target문장의 대응된 단어들이나 구들을 매치해준다. 즉 기존 통계 기반 기계번역의 TM부분을 더욱 명확하게 모델링한 방법론이다. 현존하는 고성능 기계번역 서비스는 거의 대부분 이 모델을 기초로 하고 있고 이러한 NMT 모델은 전통적인 통계 기반 기계번역의 모델링 방법과 큰틀에서 다르지 않으므로 현재 NMT는 기존 연구들의 연속이라고 생각한다.

최근에는 attention mechanism을 더욱 적극적으로 사용하는 연구가 활발히 이루어지고 있다. 기존의 방식이 source 문장 구성 요소와 이에 대응되는 target 문장 구성 요소 사이의 관계를 찾는 것에 한정되는 Inter-attention 방식이었다면, 최근 모델은 독립적으로 source, target 각 문장의 요소 내부에서 관계를 찾는 Intra-attention 까지 활용한다. 이 방식은 각 구성 성분들이 담고 있는 ambiguity의 해소에 도움을 주어 결국 번역 성능 향상으로 이어지고 점점 전통적인 신경망 모델들을 대체하고 있다[*10].

NMT에 관련해서는 따로 지면을 할애하여 설명하겠지만, 기계번역기 모델링으로 sequence-to-sequence 모델링 방법은 자연어처리를 비롯한 여러 작업(task) 뿐만 아니라, 시퀀스(sequence)를 다루는 문제들을 해결하는 한 예시의 역할을 하고 있어, 그 의미가 더욱 크다.

***1** 참고 | http://www.mt-archive.info/Weaver-1949.pdf ***2** 참고 | https://youtu.be/K-HfpsHPmvw ***3** 참고 | 1달러=1,093원 기준 ***4** 참고 | http://www.grandviewresearch.com/industry-analysis/machine-translation-market ***5** 참고 | 기업 내부의 일상적 활동 수행이 이뤄질 수 있도록 하는 기업 네트워크 시스템 ***6** 참고 | 1위안=168.23원 기준 ***7** 참고 | https://ko.wikipedia.org/wiki/구문_분석 ***8** 논문 | Sutskever, I. et al. (2014). Sequence to Sequence Learning with Neural Networks, NIPS ***9** 논문 | Bahdanau, D. et al. (2014). Neural machine translation by jointly learning to align and translate, ICLR ***10** 논문 | Vswani, A. et al. (2017). Attention Is All You Need, doi : arXiv:1706.03762