테크니컬 리포트

2019년

인공지능

학습용

데이터 구축

자연어 영역

한-영 번역 말뭉치

개요: 한-영 번역 말뭉치란?

- 번역 말뭉치

번역 말뭉치란 원문 대 번역문의 쌍을 대량으로 구축한 데이터를 말한다. 따라서 한-영 번역 말뭉치란 한국어 원문에 대응하는 영어 번역문을 쌍으로 대량으로 구축한 데이터를 말한다. 여기서 번역문은 인간번역사에 의해 번역한 결과를 말하며 최근에는 자동번역(MT, Machine Translation) 시스템에 의해 자동 번역한 결과를 인간번역사가 후편집(PE, Post-Editing)한 결과도 번역문으로 간주하는 추세이다. 한-영 번역 말뭉치의 기본적인 사례를 보이면 다음과 같다.

원문	번역문
이번 신제품 출시에 대한 시장의 반응은 어떤가요?	How is the market's reaction to the newly released product?
판매량이 지난번 제품보다 빠르게 늘고 있습니다.	The sales increase is faster than the previous product.
그렇다면 공장에 연락해서 주문량을 더 늘려야겠네요.	Then, we'll have to call the manufacturer and increase the volume of orders.
네, 제가 연락해서 주문량을 2배로 늘리겠습니다.	Sure, I'll make a call and double the volume of orders.
지난 회의 마지막에 논의했던 안건을 다시 볼까요?	Shall we take a look at the issues we discussed by the end of the last meeting?
그보다는 이번 주 새로운 주제가 더 급한 것 같습니다.	I believe that this week's new issues are more urgent.
그럼 새로운 안건으로 회의를 시작하도록 하죠.	Then, let's begin our meeting with the new issues.
네, 자료는 여러분의 앞에 미리 준비되어 있습니다.	Sure, the related materials are ready in front of you.
이번 주 금요일까지 2천개를 더 주문하라는 건가요?	Do you mean we need to order additional 2,000 items by this Friday?
네, 시간이 조금 촉박하기는 하지만 가능해 보이는데요.	Yes, time is running short, but we can manage it.

그림 1 한-영 번역 말뭉치의 사례

- 한-영 번역 말뭉치의 구축 및 활용 목적

한-영 번역 말뭉치를 구축하는 목적은 한국어 원문에 대응되는 영어 번역문을 정답(Reference)으로 활용하기 위한 것이다. 좀더 구체적으로 한-영 번역 말뭉치의 활용 목적을 나누면 세가지로 나눌 수 있다. 첫번째는 번역가 지원(CAT, Computer-Assisted Translation) 도구에서 활용할 수 있다. CAT 도구에 번역 말뭉치를 번역 메모리(TM, Translation Memory)로 활용함으로써 입력된 한국어 원문에 가장 유사한 영어 번역문을 유사도(Similarity) 계산을 통해 제공함으로써 기존에 번역된 번역문을 재활용하는 것이다. 두번째는 MT 시스템에서 한국어 원문에 대응되는 영어 번역문을 생성하기 위해 학습(Learning)을 하기 위한 것이다. 통계기반 자동번역(SMT, Statistical Machine Translation) 시스템에서는 한국어 원문의 구에 대응되는 영어 번역문의 통계적인 구테이블(Phrase Table)을 만들기 위한 것이며, 신경망 자동번역(NMT, Neural Machine Translation) 시스템에서는 딥러닝(Deep Learning)을 위한 학습 데이터로 활용하기 위한 것이다. 세번째는 자동번역을 응용할 수 있는 한영 자동 통역(ST, Speech Translation) 시스템이나 다국어 정보 검색(MIR, Multilingual Information Retrieval)시스템, 다국어 화상회의, 그리고 번역을 자동으로 평가할 수 있는 번역 자동 평가 시스템 등에서 한-영 번역 말뭉치를 활용할 수 있다.

데이터셋의 구성

한-영 번역 말뭉치의 기본적인 구성은 한국어 원문에 대응되는 영어 번역문의 쌍이다. 그리고 이 데이터셋의 내용을 설명하기 위한 다양한 태그 정보들을 부착할 수 있다. 이러한 태그 정보들로 1) 문서번호, 2) 문단번호, 3) 문장번호, 4) 도메인(분야), 5) 도메인(소분야) 6) 발화자, 7) 패러프래즈 등으로 구성될 수 있다.

원문	번역문	태그 정보						
		문서번호	문단번호	문장번호	도메인(대분야)	도메인(소분야)	발화자	패러프래즈

- 문서번호: 원문이 수집된 문서의 고유 ID에 대한 정보. 예) 문서의 제목
- 문단번호: 문서의 문단별 정보나 대화 상황별 정보. 예) 문단별 일련번호
- 문장번호: 문장의 일련번호. 예) 00000001
- 도메인(대분야): 해당 문장이 속하는 분야에 대한 대분류 정보. 예) 강연, 뉴스, 토론 등
- 도메인(소분야): 도메인(대분야)을 세분화하기 위한 전문분야 정보. 예) 강연-인공지능, 뉴스-IT
- 발화자: 발화자 수에 대한 정보. 예) A-1, B-1, A-2, B-2
- 패러프래즈: 하나의 원문에 대해 여러 인간번역사가 번역한 경우, 번역문과는 다른 번역.

데이터셋의 설계 기준과 분포

데이터셋을 설계할 때 가장 중요하게 고려해야 하는 점은 데이터 밸런스이다. 한-영 번역 말뭉치를 구축하는데 비용과 시간이 많이 들기 때문에 정해진 비용 내에서 응용시스템의 효능을 최적화 하도록 노력하여야 한다. 그러기 위해서는 해당 분류기준에 따라 골고루 데이터가 분포되도록 설계하여 학습 시 예상할 수 있는 데이터 편향성을 최소화 하는게 중요하다. 따라서 한-영 번역 말뭉치를 골고루 분포시키기 위한 설계 기준으로 도메인(대분야)와 도메인(소분야)가 중요하다.

도메인(대분야)와 도메인(소분야)의 구성 원칙과 주요 특징은 다음과 같다.

- 도메인(대분야): 도메인(대분야)는 문체와 연관되어 있다. 문체는 2명 이상이 대화하는 식의 대화체와 1명이 기술하는 식의 서술체로 구분할 수 있으며, 서술체는 말하는 식의 구어체와 글을 쓰는 식의 문어체로 구분할 수 있다. 현재 대다수의 번역 말뭉치 구축량은 문어체 > 구어체 > 대화체 순으로 구축되어 있는 실정으로 문어체는 자동번역에, 구어체는 자동통역에, 대화체는 다국어 화상회의 등의 응용시스템의 학습데이터로 활용될 수 있다. 이와 같은 문체를 기반으로 도메인(대분야)은

전문분야를 통합할 수 있는 기능적인 명칭을 의미하며, 강연, 뉴스, 토론, 문서(논문) 등으로 나눌 수 있다. 이상을 요약하면 다음과 같다.

	서술체		대화체
	문어체	구어체	
도메인(대분야)	뉴스, 문서(논문)	강연	토론

- 도메인(소분야): 도메인(대분야)를 세분화하기 위해서 통상적으로 사용하는 전문분야로 세분화한 것을 말한다. 예를 들어, 도메인(대분야)인 '강연'에 대한 전문분야는 한-영 병렬말뭉치인 경우는 K-MOOC의 분류체계를 도메인(소분야)로 활용할 수 있으며, 영-한 병렬말뭉치인 경우는 TED-Topics의 분류체계를 도메인(소분야)로 활용할 수 있다. K-MOOC나 TED-Topics에서는 구축된 강연동영상들의 분포를 매년 제공하고 있기 때문에 도메인(소분야)별로 균형있는 원문을 수집하는데 참고할 수 있다. 그리고 수집된 원문은 단어 길이나 어절 길이에 따라 원문의 길이에 따라 그 분포를 조절할 수 있다.

❑ K-MOOC 강연 분류(2017.4.7현재)

대분류				
인문	55	23%	언어 · 문학 21 인문 34
사회	70	29%	경영 · 경제 38 법률 4 사회 28
교육	10	4%	교육일반 8 특수교육 2
공학	45	19%	건축 3 토목 · 도시 1 교통 · 운송 1 기계 · 금속 12 전기 · 전자 9 정밀 · 에너지 1 컴퓨터 · 통신 14 산업 3 화공 1
자연	43	18%	생물 · 화학 · 환경 20 생활과학 5 수학 · 물리 · 천문 · 지리
의약	6	2%	의료 5 치료 · 보건 1
예체능	12	5%	응용예술 2 무용 · 체육 3 미술 · 조형 4 연극 · 영화 2 음악 1
계	241	100%	

❑ TED-Topics 강연 분류 (2017.4.7현재)

Popular topics			
Technology	695	25%
Entertainment	295	11%
Design	404	15%
Business	336	12%
Science	533	19%
Global issues	486	18%
계	2,749	100%

그림 2 도메인(소분류)를 위한 K-MOOC과 TED-Topics의 강연 분포

데이터 구조

데이터셋에 따른 항목과 해당 값은 아래 테이블과 같다.

원문	번역문	태그 정보						
		문서번호	문단번호	문장번호	도메인 (대분야)	도메인 (소분야)	발화자	패러프래즈
출발언어의 문장	목표언어의 문장	문서제목	문서의 chapter명	문서의 문장 일련번호	강연, 토론, 뉴스, 문서(논문)		A, B, C,...	번역문과 다른 다수의 번역자의 번역문

데이터 예시

이 데이터는 한-영 번역 말뭉치를 엑셀에서 구축한 예시로서, 한국어 강연을 예시로 하였다.

원문	번역문	문서번호	문단번호	문장번호	도메인(대분야)	도메인(소분야)	발화자	패러프래즈
저는 너무 기쁘기도 했고 선생님에게 칭찬을 들어서도 또 기뻤습니다.	I was so pleased and glad to hear a compliment from my teacher.	M1	1	1	강연	예체능	A	
시간이 잠깐 흐른 뒤에 저는 작은 깨달음을 얻었어요.	After a short period of time, I learned a lesson.	M1	1	2	강연	예체능	A	
처음에 요가 동작이 안 됐잖아요?	At first, I couldn't make yoga moves.	M1	1	3	강연	예체능	A	
그러면 끝까지 동작이 안 되야 되는데 동작이 된 거에요.	I thought it would continue, but somehow I made the moves.	M1	1	4	강연	예체능	A	
그리고 저에게 작은 질문을 했습니다.	So I asked myself.	M1	1	5	강연	예체능	A	
이건 뭘까? 라고.	"What is this situation?"	M1	1	6	강연	예체능	A	
다시 답이 돌아 왔어요.	Again, the answer came back.	M1	1	7	강연	예체능	A	
저는 늘 항상 열심히 했다고 생각했는데 한계를 정해 놓고 있던 거였습니다.	I thought I did my best all the time, yet I set a limit to myself in reality.	M1	1	8	강연	예체능	A	
그 한계를 정해 놓았기 때문에 저는 성공할 수가 없었어요.	For that limit, I wasn't able to succeed.	M1	1	9	강연	예체능	A	
그리고 제가 살아온 삶을 뒤돌아 보기 시작했습니다.	So I looked back on how I had lived.	M1	1	10	강연	예체능	A	

데이터 구축 과정

한-영 번역 말뭉치는 한국어 원문을 수집하고 번역자를 모집하여 올바른 영어 번역문을 구축하는 절차로 구축된다. 각각의 데이터 구축 과정을 그림으로 나타내면 다음과 같다.

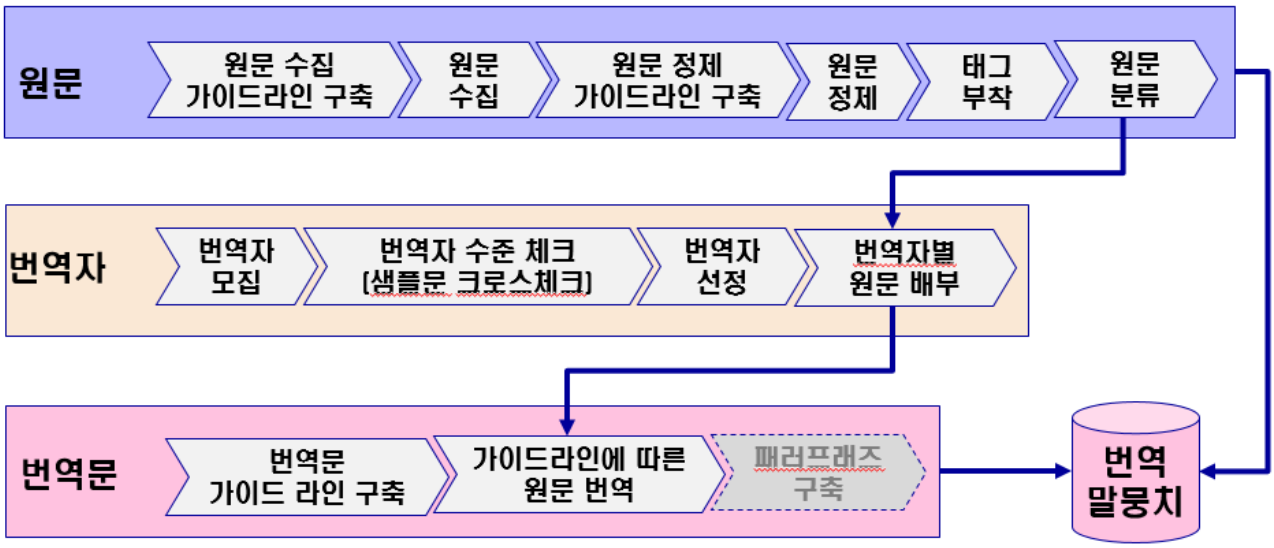

그림 3 번역 말뭉치 구축 절차

- 원문

  - 원문 수집 가이드라인 구축: 원문을 수집하기 위한 가이드라인을 만드는 단계로써, 원문의 분야, 원문의 문체, 원문 분야에 따른 양, 크라우드소싱(Crowsourcing) 또는 크롤링(Crawling)으로 수집할지의 수집 방법 등을 정하는 단계이다.

  - 원문 수집: 원문 수집 가이드라인에 따라 원문을 수집하는 단계이다.

  - 원문 정제 가이드라인 구축: 수집된 원문에 1) 오류가 있을 때 수정하는 방법, 2) 원문이 문장 단위가 아니거나 문장 단위를 넘어설 때 수정하는 방법, 3) 원문에 개인정보가 있을 때 일반화하는 방법, 4) 특수 기호로 원문이 깨질 경우 수정하는 방법 등과 같은 가이드라인을 정하는 단계이다.

  - 원문 정제: 원문 정제 가이드라인에 따라 원문을 정제하는 단계이다.

  - 태그 부착: 원문 수집 가이드라인에 따라 설정한 원문의 1) 문서번호, 2) 문단번호, 3) 문장번호, 4) 도메인(대분야), 5) 도메인(소분야), 6) 발화자, 7) 패러프래즈 등과 같은 원문의 메타정보를 부착하는 단계이다.

  - 원문 분류: 밸런스 있는 번역 말뭉치를 구축하기 위한 단계로써 부착된 태그에 따라 원문의 양을 분류하는 단계이다.

- 번역자

  - 번역자 모집: 원문의 언어와 번역할 언어별로 인간번역사를 모집하는 단계이다.

- 번역자 수준 체크(샘플문 크로스 체크): 번역할 번역자들의 수준을 체크하는 단계로써, 샘플문에 의한 번역을 실시하고 번역자간 크로스 체크(Cross-check)를 실시하여 적정 수준의 번역 품질을 보장할 수 없는 번역자를 거르는 단계이다.
- 번역자 선정: 번역자 수준 체크 단계에 의해 번역자를 번역량에 따라 최종적으로 번역자를 선정하는 단계이다.
- 번역자별 원문 분배: 납품 단계에 따라 번역할 원문의 양을 나누어 번역자별로 원문을 분배하는 단계이다.

- 번역문

- 번역 가이드 라인 구축: 원문을 번역문으로 만드는 데 있어 번역자가 참고해야 하는 번역 가이드라인을 구축하는 단계로써, 한-영 번역 말뭉치의 경우 1) 의역의 수준, 2) 인명 번역, 3) 기호 번역, 4) 약어 번역, 5) 문체 번역, 6) 문맥의 반영 정도, 7) 후편집(Post-editing) 방법 등에 대한 가이드라인을 구축해야 한다.
- 가이드라인에 따른 원문 번역: 번역 가이드 라인에 따라 원문을 번역하는 단계로써, 후편집(Post-editing)의 양, 언어별 구축량 등을 정해야 한다.
- 패러프래즈 구축: 원문에 대응되는 번역문을 1개로 할 지, N개로 할 지를 결정하는 단계로써, 원문에 대응되는 다수의 번역문을 구축할 경우 향후 자동번역 결과의 자동 평가에 활용할 수 있다. 그러나 패러프래즈 구축은 필수적으로 필요한 것은 아니다.

검수와 품질 확보

한-영 번역 말뭉치를 구축할 때 대량으로 구축하는 것도 중요하지만, 구축된 번역 말뭉치의 품질을 확보하는 것이 더욱 중요하다. 구축된 한-영 번역 말뭉치의 모든 문장에 대해 전수 검수를 실시하는 것이 가장 좋겠으나, 구축양이 대량이므로 일정한 구축량에 대해 샘플 검수를 실시하여 적정한 품질이 확보될 때까지 반복적으로 검수를 실시하여 품질을 확보하는 것이 바람직하다. 한-영 번역 말뭉치의 적정한 품질을 확보하기 위한 품질 검수 절차는 다음과 같이 할 수 있다.

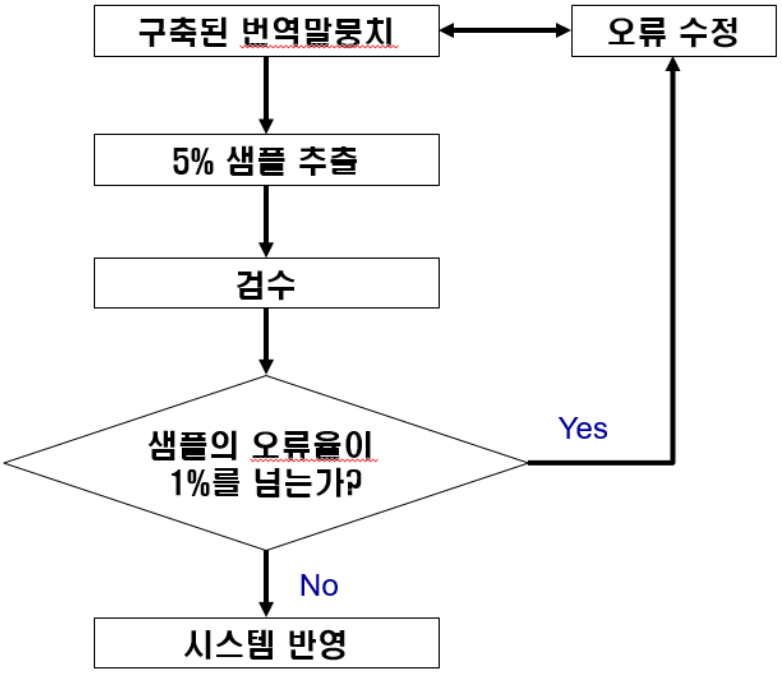

그림 4 품질 확보를 위한 품질 검수 절차

예를 들어, 한-영 번역 말뭉치가 대상이라고 할 때, 한-영 번역 말뭉치가 구축되면 구축된 말뭉치 중에 5%를 샘플로 추출하여 구축 가이드라인에 맞게 한-영 번역 말뭉치가 구축되었는지를 검수한다. 검수 결과, 5%의 샘플 중에 오류율이 1%를 넘기지 않으면 해당 시스템에 반영하고 나머지 한-영 번역 말뭉치를 대상으로 상기의 절차를 반복한다. 검수 결과, 5%의 샘플 중에 오류율이 1%를 넘으면 구축된 한-영 번역 말뭉치를 대상으로 오류 수정을 전체적으로 실시하고 5% 샘플을 새로 추출하여 상기의 절차를 반복하여 적정 품질을 확보하도록 한다.

데이터 구축 담당자

최승권 (전화: 042-860-6909, 이메일: choisk@etri.re.kr)