관용구 기계번역을 위한 한-영 데이터셋 구축 및 평가 방법

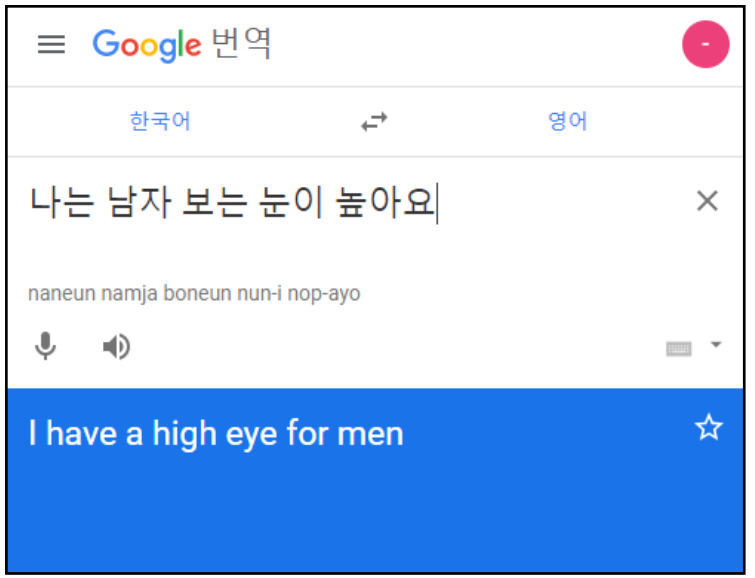
최민주

mjchoi0831@gmail.com

1. 서 론

- NMT 기반 기계번역은 훌륭한 번역 성능을 보이나, 종종 오역이 발생함
  - 예) 관용구 번역

- 기계번역은 함축적인 의미를 지닌 관용구를 정확하게 번역/평가할 수 없음
  - 관용구 번역에 특화된 데이터셋과 평가방법이 필요



1. 서 론

- NMT에서 관용구를 학습/평가하려면 뭐가 필요할까
  - 관용구가 포함된 다량의 문장 **데이터셋**
    - 한-영 번역 쌍 말뭉치로부터 관용구 포함 번역 쌍 추출

      ⇒ 관용구 학습 데이터셋 'KISS' 구축

  - 관용구 번역에 특화된 번역 **평가 지표**
    - 관용구 학습 데이터셋을 이용

      ⇒ 블랙리스트(평가 지표) 생성

- KISS : **K**orean-english **I**dioms in **S**entences data**S**et
- https://github.com/Judy-Choi/KISS

3. KISS : 관용구 포함 한-영 번역 쌍 데이터셋 구축

표준국어대사전
수록 관용어

...

눈이 높다

마침표를 찍다

총대를 메다

화촉을 밝히다

...

AI Hub 한국어-영어 번역(병렬) 말뭉치

KO	EN
나는 여자 보는 눈이 높아요	I have **high standards** for woman.
아버지의 명예회복을 위한 김지훈의 기나긴 여정이 마침내 마침표를 찍었다.	Kim Ji-hoon's long journey ... has finally **come to an end**.
기획재정부가 혁신성장 관련 규제 완화에 총대를 멘다.	The Ministry of Economy and Finance **takes charge of** ...
중국 출신 할리우드 스타 장쯔이 ...와 내년 화촉을 밝힌다.	Chinese Hollywood star Zhang Ziyi (28) will **marry** ...
...	...

3. KISS : 관용구 포함 한-영 번역 쌍 데이터셋 구축

- 관용구 수집
  - 표준국어대사전 온라인 사이트 → 3,887개 관용구 목록 다운로드
- 한-영 문장 번역 쌍 추출
  - AI Hub 한국어-영어 번역(병렬) 말뭉치
    - 430개 한국어 관용구 포함 18,808개 번역 쌍 추출
  - 동일한 관용구 포함 문장 4~40개로 제한
    - 420개 한국어 관용구 포함 7500개 번역 쌍 추출 => KISS 구축

- **KISS** : **K**orean-english **I**dioms in **S**entences data**S**et

github.com/Judy-Choi/KISS

Problem...

- 구축한 데이터셋에 다량의 오역 쌍 존재
  - 데이터셋에 포함된 다량의 오역을 제거할 수 있는 방법 필요
- 관용구 번역 시 다량의 오역 발생
  - 기계번역 오류를 탐지할 수 있는 품질 평가 지표 필요

관용구	운을 떼다	유명을 달리하다
한국어 원문	정부가 노인 연령 기준을 높이는 방안에 대해 운을 뗐다.	오키나와 현지사인 오나가 다케시가 지난 8일 췌장암으로 유명을 달리했다.
영어 번역 쌍	The government has been **lucky** about ways to raise the criteria of age for senior citizens.	Takeshi Onaga, an incumbent Governor of Okinawa, became **famous** for pancreatic cancer on the 8th.

4. 블랙리스트 평가 방법

- 직역으로 인한 번역 오류 탐지
- 단어-대-단어 구조의 관용구를 번역한 결과로부터 오역 여부를 판별
- 원리
  - 관용구를 한-영 번역한 결과에 블랙리스트 단어가 1개 이상 포함되어 있으면 오역으로 간주

관용구	블랙리스트
꼬집어 말하다	nip pinch twitch
눈 높다	eye
운을 떼다	lucky
유명을 달리하다	famous

관용구	눈이 높다
블랙리스트	**eye**
한국어 원문	나는 여자 보는 눈이 높아요.
영어 번역 쌍	I have high **eye** for woman.

블랙리스트를 이용한 오역 탐지

5. 평가

- 블랙리스트 구축
  - 420개 관용구 중 275개 관용구에 대한 블랙리스트 구축
    - 오역이 거의 없는 관용구 제외
      - 예) '그건 그렇고' → 'By the way' 로 대부분 정확하게 번역됨
    - 영어로 직역되는 관용구 제외
      - 예) '(사람의 마음을) 가지고 놀다'→ 'play with' 로 직역되므로 블랙리스트 구축 불가
- 블랙리스트를 이용한 오역 없는 번역 쌍 추출
  - KISS 로부터 275개 관용구 포함 3,461 개 번역 쌍 추출

5. 평가

- 오역 제외한 번역 쌍을 이용한 기계번역 서비스 품질 평가
  - 관용구 275개
  - 번역 쌍 3,461개

	Google 번역	Naver Papago	Kakao i 번역
블랙리스트 탐지	1,179	1,093	1,049
블랙리스트 미탐지	2,282	2,368	2,412
* 번역 정확도 (%)	65.93	68.41	**69.69**
평균 BLEU 점수	30.04	13.47	33.83

* 번역 정확도 : 전체 문장 중 블랙리스트 미탐지된 문장 수

5. 평가

- 번역 품질과 BLEU 점수가 비례하지 않는 예시

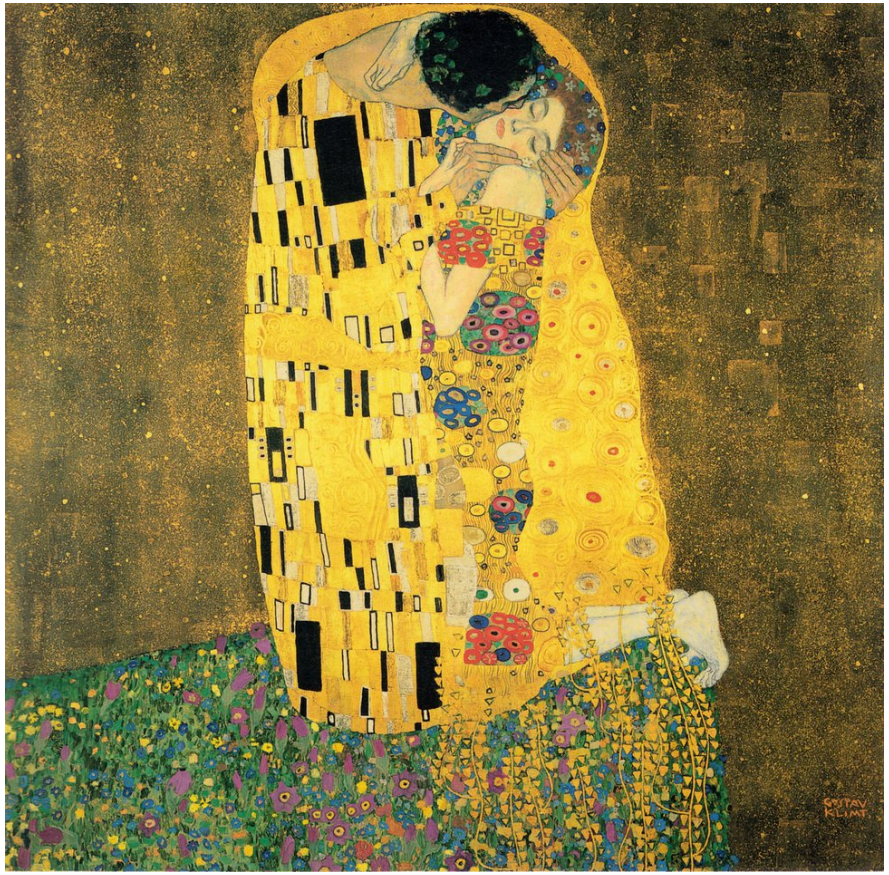
관용구	가닥을 잡다
한국어 원문	당초 강 위원장은 국정감사에서 이 문제가 불거지자 "감사원 감사를 받겠다" 며 버텼지만 즉각 사퇴로 **가닥을 잡았다**.
영어 번역 쌍	At the beginning, Representative Kang said, "I will be audited by the auditor," when the matter was raised in the state audit.
Naver Papago 번역 결과	Initially, Kang endured the issue during a parliamentary audit, saying he would undergo an audit by the Board of Audit and Inspection, but he immediately **decided to** step down.
BLEU 점수	**4.7930e-76**

6. 결론

- 관용구가 포함된 다량의 문장 **데이터셋 구축**
  - KISS : 420개 관용구, 7,500개 한-영 번역 쌍 데이터셋

- 관용구 번역에 특화된 번역 **평가 지표 생성**
  - 275개 관용구, 3,461개 번역 쌍에 대한 블랙리스트

- 블랙리스트를 이용한 기계번역 서비스의 번역 정확도 측정 방법 제안



감사합니다 – Thank you