Multilingual Machine Translations

Ki Hyun Kim

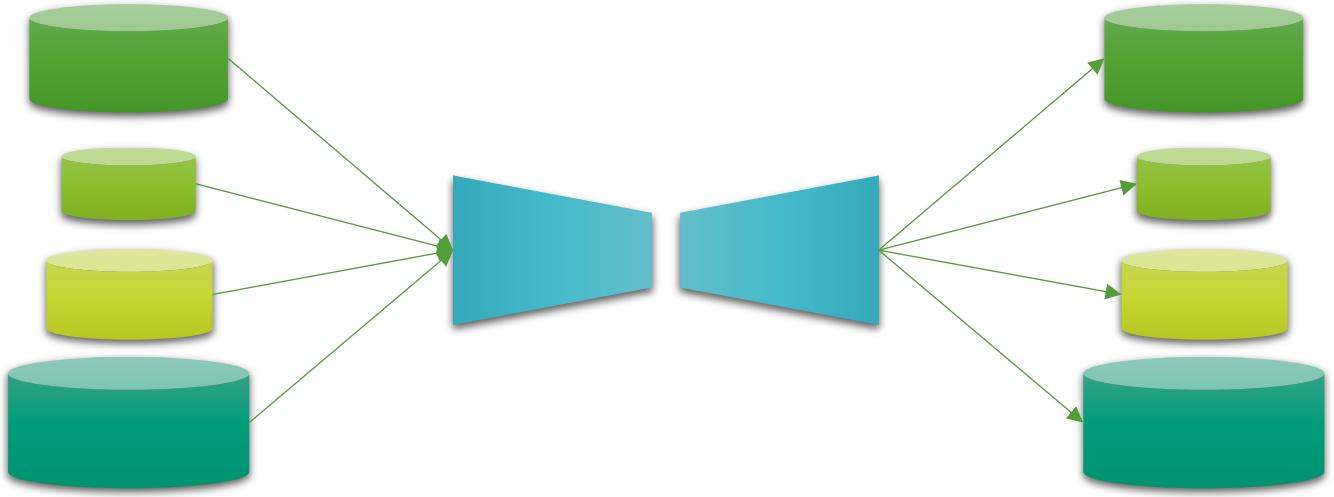
nlp.with.deep.learning@gmail.com

Motivations

- 병렬 코퍼스 당 모델 2개가 나옴
  - e.g. 한영 코퍼스 → 한영 번역 모델, 영한 번역 모델

- 언어 쌍에 따른 데이터 양의 편차가 심함
  - e.g. 한국어/태국어 번역?

- 하나의 모델로 모든 언어 쌍을 커버하자 – Multi-task Learning



Zero-Shot Translation [Johnson and Schuster et al., 2017]

- Previous
  - Hello, how are you? → Hola, ¿cómo estás?
- Proposed
  - <2es> Hello, how are you? → Hola, ¿cómo estás?

- Many to One
  - 다수의 언어를 encoder에 넣고 훈련
- One to Many
  - 다수의 언어를 decoder에 넣고 훈련
- Many to Many
  - 다수의 언어를 encoder와 decoder에 모두 넣고 훈련
- Zero-shot Translation
  - 위의 방법으로 훈련된 모델에서 zero-shot translation의 성능을 평가

In Practice,

- 굳이 하나의 모델로 모든 언어 쌍을 cover할 이유가 없다.

- 오히려 domain adaptation에 활용 가능
  - 전문분야 도메인
    - <2general>, <2medical>, <2legal>
  - 어투
    - <2polite> vs <2rude>
    - <2literature> vs <2converstion>

Conclusion

- 여러 병렬 코퍼스를 한 모델에 학습하게 되면,
  - 비슷한 언어적 특성을 갖는 언어 쌍의 번역 성능에 서로 도움을 받을 수 있다.

- 입력 시퀀스의 첫 번째 토큰에 조건(e.g. <2es>)을 위한 특수 토큰을 넣어 학습
  - 실무에서는 오히려 domain adaptation에 적용할 수 있는 방법