Appendix:
Align Parallel Corpus

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Why we need

- Parallel Corpus?

English	Korean
I love to go to school.	나는 학교에 가는 것을 좋아한다.
I am a doctor.	나는 의사 입니다.

- 주요 수집 대상
  - 뉴스 기사, 드라마/영화 자막

- 대부분의 경우, 문서 단위의 matching은 되어 있지만, 문장 단위는 되어 있지 않음

Champollion

- 최초로 이집트 상형문자를 해독한 역사학자

(Jean-François Champollion, Image from Wikipedia)

- Perl로 제작 됨
  - https://github.com/LowResourceLanguages/champollion
- ratio parameter의 역할
  - source 언어의 character 당 target 언어의 character 비율
- 결과 형태:

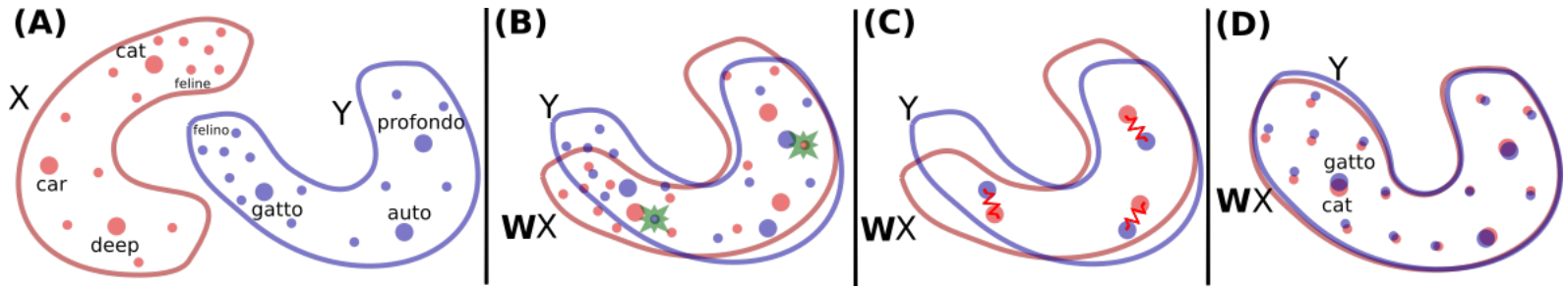
```
omitted <=> 1
omitted <=> 2
omitted <=> 3
1 <=> 4
2 <=> 5
3 <=> 6
4,5 <=> 7
6 <=> 8
7 <=> 9
8 <=> 10
9 <=> omitted
```

단어 번역 사전에 기반하여
사전을 최대한 만족하는 sentence align을 찾는 방식

How to build Word Translation Dictionary

- MUSE: https://github.com/facebookresearch/MUSE

- This project includes two ways to obtain cross-lingual word embeddings:
  - **Supervised**: using a train bilingual dictionary, learn a mapping from the source to th e target space.
  - **Unsupervised**: without any parallel data or anchor point, learn a mapping from the s ource to the target space.



Example: MUSE Result

```
stories <> 이야기
stories <> 소설
contact <> 연락
contact <> 연락처
contact <> 접촉
green <> 녹색
green <> 초록색
green <> 빨간색
dark <> 어두운
dark <> 어둠
dark <> 짙
song <> 노래
song <> 곡
song <> 음악
salt <> 소금
```

Procedure to Build Parallel Corpus

1) Bi-lingual Corpus 정제 (e.g. 노이즈 제거)
2) Tokenization 수행 (No subword segmentation)
3) 각 언어별 코퍼스에 대해서 word embedding 수행 (FastText 활용)
4) MUSE를 활용하여 word translation dictionary 추출
5) Champollion을 활용하여 align 수행