Word Embedding Evaluation

Reference

- Evaluation methods for unsupervised word embeddings, Tobias Schnabel et al.

단어 임베딩 종류는 word2vec, glove, fasttext 등의 prediction based 기법과 PMI, SVD를 이용한 count-based 기법이 있다. 여러 개의 단어 임베딩을 만들었을 때, 어떤 임베딩의 품질이 좋은지 판단을 해야한다.

물론 machine learning task에서 text를 vector로 embedding하고, 이를 input으로 넣어 ML의 성능을 비교함으로써 단어 임베딩의 품질을 비교해볼 수 있다. 그러나 ML의 성능을 내기까지 hyperparameter tuning, train, test의 일련의 과정을 거쳐야하며 이는 많은 시간을 필요로 한다. 따라서 여기서는 임베딩 벡터 자체만으로 할 수 있는 평가를 살펴본다. 결국 어떤 task에 대한 정확도 비교이긴 한데, ML을 이용한 비교보다는 시간이 훨씬 덜 걸릴 것이다.

Absolute intinsic evaluation

임베딩 벡터가 각각 평가되며 그들의 최종 점수가 비교된다. 총 네 개가 있다.

- Relatedness: 어떤 두 단어 A, B에 대해서 사람이 매긴 점수와 두 단어 A, B의 임베딩 벡터 간의 코사인 유사도와의 유사성을 본다. 여기서 유사성은 correlation으로 정의되며 코사인 ㅇ사도와 사람이 매긴 점수가 높은 correlation을 가지면 임베딩 벡터의 품질이 좋은 것이다.
- Analogy: Mikolov가 word2vec과 함께 제안한 방법이다. x : y의 관계를 가지는 a : b를 찾는다. 예를 들어 한국 : 서울 = 일본 : 도쿄 이런 식이다. Analogy test에는 3CosAdd와 3CosMul 방법이 있다.
- Categorization: 다른 카테고리로 단어의 clustering을 복구하는 것이 목표이다. 이를 위해 임베딩 벡터들이 군집화되고 labeled dataset에 대해서 반환된 cluster의 purity가 계산된다.
- Selectional preference: 명사가 동사와 맺는 관계가 주어인지, 목적어인지 밝히는 것이 목적이다. 예를 들어 people eat이라고 하지 eat people라고는 말하지 않는다.

Dataset for similaity test (for English)

- WordSim-353 dataset
- MEN dataset
- Turk dataset
- Rare words dataset
- SimLex-999 dataset

Dataset for Analogy test (for English)

- MSR dataset (아직 못 찾음)
- Google Analogy dataset

Comparative intrinsic evaluation

이 평가 방법은 사람들에게 선호하는 임베딩을 물어보는 것이라고 한다. 논문에서 정확히 무엇을 뜻하는 것인지 모르겠으므로 패스.

Extrinsic Tasks

위에서 언급했었는데, 임베딩 벡터가 특정 task에 미치는 정도를 측정한다. 품질 좋은 임베딩 벡터가 downstream task의 결과 향상을 이끈다고 가정한다. 하지만 본 논문에서는 다른 task는 서로 다른 임베딩 벡터를 선호함을 알아냈다. 즉, 다른 task에 대해 특정 임베딩 벡터만 가장 좋은 성능을 내지는 않는다는 것이다. 따라서 본 논문에서는 이런 extrinsic task을 임베딩 벡터의 품질을 알아내는 방법으로 사용하지 않을 것을 권고한다.