소프트웨어종합설계 최종보고서

Effective ways to Select a Dataset from Large Corpus

input.txt

팀원
김주찬
김유진
Dobreva Iva

지도교수	여진영
조교	이가연

Contents

1. 연구의 주제

자연어 처리 분야에서는 학습에 방대한 양의 데이터를 사용하게 되는데, 데이터 수의 증가가 반드시 학습 성능의 향상으로 이어지지는 않는다. 따라서, 성능을 유지하면서 데이터의 개수를 줄여 학습의 효율성을 높이는 방법이 필요하다고 할 수 있다. 우리는 이런 상황 속에서, 그래프 형태의 데이터에 영향력 확산 문제를 적용하여, 데이터를 효율적으로 선택하는 방법에 대해 연구했다.

2. 연구의 필요성

1) 기존 연구

학습에 방대한 양의 데이터를 사용하게 되어 증가한 학습 비용을 줄이기 위한 방법들로, Network Architecture Modification, Transfer Learning, Active Learning, Data Selection 등이 있다. 모델의 구조 자체를 압축하는 Network Architecture Modification과, 이미 학습된 모델의 데이터를 이용하는 Transfer Learning은 모델 관점에서 접근한 방법이다. 모델이 학습 과정에서 학습할 데이터를 고르게 하는 Active Learning이나 학습할 데이터를 학습 이전에 고르는 Data Selection은 데이터 관점에서 학습 비용을 줄이고자 한 것이다.

2) 연구의 필요성

Data Selection의 목표는 커다란 양의 데이터 중에서 학습 성능을 향상시킬 수 있는 좋은 데이터를 선택하는 것이다. 그러나 기존 자연어 처리 연구에서 사용하는 데이터는 대부분 단순한 리스트 형태이며, 텍스트 파일 형식으로 저장된다. 이러한 한계에서 벗어나고자, 이 연구에서는 그래프 형식의 데이터를 이용하는 새로운 접근을 시도해보려 한다. 그래프 형식의 데이터에 영향력 최대화 문제를 적용하여 학습에 사용할 데이터를 선택한다면, 학습 성능의 저하를 최소화하여 학습 비용을 줄일 수 있을 것으로 기대된다.

3. 연구 방향

1) 학습 모델 선정

- ESIM -> CEDR

처음 사용하기로 했던 모델은 ESIM으로, context와 response로 이루어진 대화 데이터를 사용해 여러 개의 response 중에 적절한 것을 고르는 문제를 푸는데 사용되는 모델이었다. 하지만 ESIM을 사용해 학습하는 과정에서 너무 시간이 오래 걸리고, 학습 중반부부터 epoch가 증가할수록 mean average precision이 감소하거나 비슷해지는 문제가 발생했고, 이 연구의 목표는 모델 관점이 아닌 데이터 관점의 접근이므로, 용이한 실험을 위해 모델을 변경하기로 했다.

변경 후에 사용하기로 한 모델은 CEDR로, Query와 Document로 이루어진 데이터를 이용해, 어떤 query에 대해서 해당 document가 그 query와 관련이 있는가를 판단하는 문제를 푸는데 사용되는 모델이다. 이 모델의 특징은, 이전의 연구들이 맥락 정보를 충분히 반영하지 못하는 문제점을 해결하기 위해, 미리 학습된 ELMo나 BERT와 같은 모델을 PACRR, KNRM, DRMM 같은 기존의 document ranking 모델에 결합했다는 것이다. 이 연구에는 BERT와 KNRM의 결합을 사용했다.

모델이 변경되었으므로 구현해야 하는 데이터 그래프의 형태도 바뀌었는데, 이전에 그래프의 노드가 context와 response로 구성된 대화 데이터를 담고 있었다면, 이제는 document 데이터가 노드를 구성하게 되었다. 엣지는 각 노드를 구성하는 document 데이터 사이의 유사도에 따라 결정된다.

2) 연구 목표

- 그래프 구현 -> 영향력 최대화 -> 학습

기존 데이터를 그래프 형태로 바꾼 다음, 영향력 최대화 알고리즘을 이용해 영향력 확산을 최대화할 수 있는 초기 데이터셋을 고른다. 이렇게 선택된 데이터셋을 이용해 자연어 처리 모델 CEDR을 학습시킨 다음, 기존 데이터셋을 이용해 학습시킨 결과와 비교한다.

4. 연구 내용

1) 그래프 구현

① 그래프 구조

원래의 데이터는 하나의 query에 대해, 각 document가 그 query와 관련이 있는가에 대한 데이터를 갖고 있다. 이 데이터의 형태를 바꾸어, 노드는 doucument 데이터와 해당 document와 각 query들 간의 연관성 정보로 구성되도록 했다. 엣지는 데이터에 포함되어 연관성을 나타내는 rel_1d 값에 따라 결정이 되는데, rel_1d는 어떤 document가 해당 query와 관련이 있으면 1, 아니면 0 값을 가진다. 이때, 엣지의 weight은 양 노드가 모두 참여하는 쿼리에 대해 양 노드의 rel_1d 값을 모두 더한 다음 양 노드가 모두 참여하는 쿼리의 개수로 나눠서 구하며, 엣지는 weight이 0이 아닐 때 연결된다.

② 구현 방법

그래프의 구현에는 두 가지 방법을 사용해 보았다.

첫 번째 방법은, 노드를 생성하고, 각 노드들 사이의 weight을 계산해서, 연관되어 있다면 연결하는데, 이것을 각 노드 쌍에 대해 반복하는 것이다. 이 방법은 그래프를 생성하는데 7시간 가량의 긴 시간이 소요된다는 단점이 있었다.

두 번째 방법은, 노드를 생성하고 모든 엣지의 리스트를 만든 다음 weight을 계산해 연관된 노드를 한 번에 연결하는 것이다. 그러나 엣지의 개수가 너무 많아 메모리 할당에 어려움을 겪는 문제점이 있어, 최종적으로 첫 번째 방법을 사용했다.

2) 영향력 최대화 문제

① 영향력 확산 모델

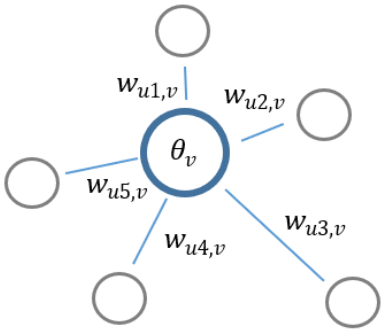
영향력 확산 모델은 소셜 네트워크에서 영향력 확산의 패턴을 분석하기 위해서 만들어진 모델이다. 영향력 확산 모델 중 가장 널리 사용되는 두 가지 유형의 모델은, independent cascade model과 linear threshold model이다.

Independent cascade model은 정보가 확산될 확률을 통해 확산을 표현하는데, k개의 초기 노드로부터 확산이 시작되어, 어떤 노드 u가 시간 t에 정보를 수용했다면 시간 t+1에 인접 노드 v에게 $p_{u,v}$의 확률로 정보를 확산시킨다.

Linear threshold model은 노드가 주변 노드들에게 받은 정보를 수용할지 여부에 의해 정보 확산이 결정되는데, 여기서 각각의 노드는 특정한 threshold를 가지고, 노드와 노드 사이의 엣지는 인접한 노드에 대한 가중치인 weight을 가지게 된다. 한 노드는 주변 노드들 중 정보를 수용한 정도가 그 threshold를 넘게 되면 정보를 수용하게 된다. 이를 식으로 표현하면, 각 노드 v에 대해 threshold를 θ_v, v와 인접한 노드 u에 대한 가중치를 $w_{u,v}$라고 할 때,

$$\sum_{u \in Neighbor\ of\ v} w_{u,v} \geq \theta_v$$

의 조건을 만족할 때 새로운 정보를 수용한다.



② 영향력 최대화 문제

이렇게 영향력 확산 모델을 정의했을 때, 영향력 최대화 (influence maximization) 문제란, 주어진 그래프 G = (V, E)에 대해, 초기에 k개의 노드가 활성화되어 있다고 가정했을 때, 영향력 함수 f(S)를 최대화하는 k개의 노드들로 이루어진 부분집합 S⊆V를 고르는 문제이다.

이러한 영향력 최대화 문제를 자연어 데이터에 적용하는 것이 이 연구의 목표이다. 주어진 데이터를 그래프 형태로 구성한 다음, 영향력 최대화 문제를 통해 영향력을 최대화할 수 있는 초기 노드를 선택하면, 자연어 처리 모델의 학습을 위한 좋은 데이터의 집합이 만들어질 것으로 기대되었다.

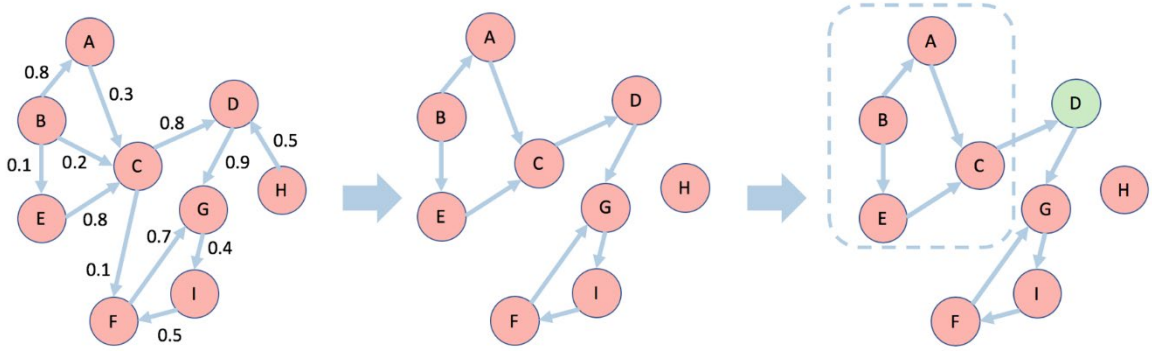
③ 영향력 최대화 알고리즘

- Greedy Algorithm

Kempe et al.은 영향력 최대화 문제의 최적 해를 구하는 것은 NP-hard라는 것을 증명했고, 따라서 영향력 함수의 submodularity 성질을 이용한 근사 알고리즘을 제안했다. S⊆V이면, 더 작은 집합 S에 임의의 노드 u를 추가한 영향력 함수의 증가량, f(S∪u)-f(S)가 적어도 더 큰 집합 T에 동일한 노드 u를 추가한 영향력 함수의 증가량, f(T∪u)-f(T) 보다 큰 것이 submodularity 성질이다. 따라서 영향력 극대화 문제를 풀기 위한 근사 알고리즘으로, 매 회마다 f(S∪u)-f(S)가 가장 큰 노드 u를 선택하는 Greedy algorithm을 제안했다. 이것은 submodularity 성질로 인해 최적해의 63%의 해를 보장한다.

그러나 greedy algorithm은 영향력 함수를 계산하는 횟수가 너무 많아, 노드와 엣지의 개수가 수천여 개에 불과한 데이터셋에 적용해도 수일이 걸릴 정도이므로, 수십만 개의 데이터를 포함하는 데이터셋에 적용하기에는 힘들다는 한계가 있었다.

- **Reverse Influence Sampling(RIS)**



RIS 알고리즘은 Reverse Reachable Set, 즉 RR set을 이용해 영향력 최대화 문제를 해결한다. 영향력 최대화 모델에서 그래프의 엣지들은 각각 확산 확률을 가지고 있는데, 이 그래프에 대해서 엣지들을 일정 확률에 의해 제거해 새로운 그래프를 생성한다. 이때 각 엣지가 제거될 확률은 1에서 확산 확률을 뺀 값이므로, 확산 확률이 높을수록 엣지가 남아 있을 확률이 높아진다. 그리고 이 그래프에서 하나의 노드 v를 대상으로, 확산 단계에서 v 이전에 있는 노드들, 즉 v로 향하는 엣지를 가진 노드들의 집합을 만드는데, 이것이 RR set이다..

그 후에, 여러 RR set들의 집합인 R을 만든 다음, R에 포함된 RR set들에서 가장 많이 나타나는 k개의 노드를 선택하게 되는데, RR set의 원소가 되는 노드는 대상 노드에 영향력을 끼칠 수 있는 노드라고 할 수 있으므로, R에서 자주 나타나는 노드는 곧 영향력이 높은 노드이다. 따라서 이 방식으로 선택한 노드는 영향력을 최대화하는 노드가 된다.

- **Two-phase Influence Maximization(TIM)**

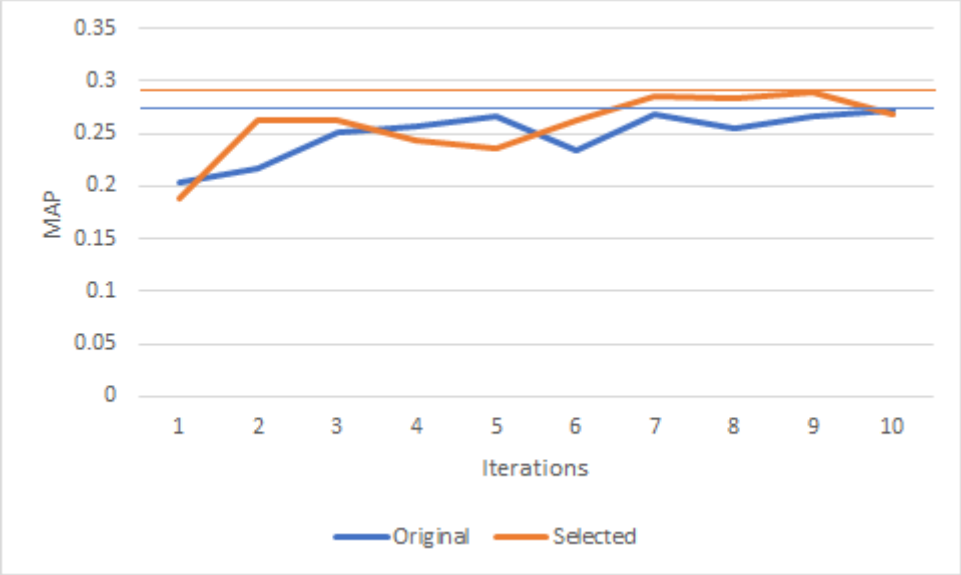
영향력 최대화에 있어서 정확성과 효율성을 둘다 만족시키려면, 얼마나 많은 RR set을 만들어야 할까에 대한 문제가 있다. RIS에서는 RR set을 생성하는데 드는 cost를 계산해서 일정 임계값을 넘으면 생성을 멈추도록 되어 있다. 여기서 더욱 효율을 높이기 위해서, Two-phase Influence Maximization, 즉 TIM 알고리즘은 RR set의 개수를 정해 놓는 방식을 사용한다. 따라서 TIM은 두 단계로 이루어지는데, 필요한 RR set의 개수를 예측하는 Parameter estimation 단계를 거친 다음 Node Selection을 하게 된다. 여기에 더욱 속도를 향상시키기 위해 parameter estimation의 성능을 높이는 중간 단계를 추가한 것이 TIM plus이다.

이 연구에서는 TIM plus를 이용했다.

3) 학습

영향력 최대화 알고리즘 TIM plus를 이용해 데이터를 선택한 다음, CEDR(Contextualized Embeddings for Document Ranking)의 학습에 이 선택된 데이터를 이용했다.

5. 연구 결과



기존 약 11만 개의 데이터에서 5만 여개의 데이터를 선택해서 학습시킨 결과이다. 데이터의 개수가 줄었기 때문에 학습에 걸린 시간도 11시간에서 6시간 정도로 줄었지만 Mean Average Precision 값을 비교해 보면 학습의 정확도는 비슷하거나 향상된 것을 확인할 수 있다. 학습 시간은 줄고 정확도는 향상되었으므로 효율적인 데이터 선택이라는 목표를 달성했다고 할 수 있다.

6. 참고 문헌

D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 137–146, 2003.

Vishal Kaushal, Rishabh Iyer, Suraj Kothawade, Rohan Mahadev, Khoshrav Doctor, and Ganesh Ramakrishnan. Learning from less data: A unified data subset selection and active learning framework for computer vision. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1289–1299. IEEE, 2019.

S. MacAvaney, A. Yates, A. Cohan, and N. Goharian. CEDR: Contextualized embeddings for document ranking. In SIGIR, pages 1101–1104, 2019.

Y. Tang, X. Xiao, and Y. Shi. Influence maximization: near-optimal time complexity meets practical efficiency. In SIGMOD 2014.

Suman Banerjee, Mamata Jenamani, and Dilip Kumar Pratihar. 2018. A Survey on Influence Maximization in a Social Network. arXiv preprint arXiv:1808.05502 (2018).