마르코프 가정(Markov Assumption)은 n-gram 언어 모델 등에서 자주 사용되는 확률적 가정입니다. 이 가정은 "현재 상태(단어)는 오직 바로 앞의 n-1개 상태(단어)에만 의존한다"는 내용입니다. 즉, 더 이전의 정보는 무시하고, 최근의 정보만을 사용하여 다음 단어의 확률을 예측합니다.

왜 마르코프 가정을 하는가?

- **계산의 단순화**: 자연어는 매우 긴 문맥을 가질 수 있지만, 모든 이전 단어를 고려하면 계산량이 기하급수적으로 증가합니다. 마르코프 가정을 적용하면, 최근 n-1개 단어만 고려하므로 계산이 훨씬 단순해집니다.
- **데이터 sparsity(희소성) 문제 완화**: 모든 가능한 단어 조합의 확률을 추정하려면 방대한 데이터가 필요합니다. 마르코프 가정을 통해 필요한 데이터의 양을 줄일 수 있습니다.
- **모델 학습 및 구현 용이**: 실제로 n-gram 모델을 구현할 때, 마르코프 가정 덕분에 효율적으로 확률을 계산하고 저장할 수 있습니다.

근거

- **자연어의 지역성(Locality)**: 실제 언어에서도 대부분의 경우, 바로 앞의 몇 개 단어가 다음 단어에 가장 큰 영향을 미칩니다. 예를 들어, "나는 밥을" 다음에는 "먹었다"가 올 확률이 높습니다.
- **경험적 성능**: 마르코프 가정을 적용한 n-gram 모델이 실제로 많은 자연어처리 작업에서 좋은 성능을 보임.

한계

- **장기 의존성(Long-term dependency) 무시**: 문맥이 길어질수록, n-gram 모델은 중요한 정보를 놓칠 수 있습니다. 예를 들어, "비가 오면 길이 미끄럽다"에서 "비가 오면"과 "미끄럽다" 사이에 많은 단어가 끼어 있으면, n-gram 모델은 이 관계를 포착하지 못합니다.
- **차원의 저주(Curse of Dimensionality)**: n이 커질수록 필요한 파라미터(확률표)가 기하급수적으로 늘어나 데이터가 부족해집니다.
- **희소성 문제**: n이 커질수록 실제로 관측되지 않은 n-gram이 많아져 확률 추정이 어려워집니다.

리소스 및 차원의 저주

- n-gram의 n이 커질수록, 필요한 메모리와 계산량이 급격히 증가합니다. 이를 차원의 저주라고 하며, 실제로는 2-gram(빅그램)이나 3-gram(트라이그램) 정도만 실용적으로 사용합니다.
- 차원의 저주를 완화하기 위해 스무딩(smoothing) 기법(예: 라플라스 스무딩 등)을 사용합니다.

요약

마르코프 가정은 자연어처리에서 계산 효율성과 데이터 요구량을 줄이기 위해 도입된 중요한 가정입니다. 하지만 장기 의존성 포착의 한계와 차원의 저주 등 단점도 존재하므로, 최근에는 RNN, LSTM, Transformer 등 더 복잡한 모델이 등장하게 되었습니다.