FastText란 무엇인가?

FastText는 Facebook AI Research(FAIR)에서 개발한 단어 임베딩(word embedding) 및 텍스트 분류(text classification) 모델입니다. Word2Vec과 유사하게 단어를 벡터로 변환하지만, **단어 내부의 subword(부분단어, n-gram)** 정보를 활용한다는 점에서 차별점이 있습니다.

FastText의 주요 특징 및 장점

1. **오타 및 희귀 단어 처리에 강함**
   - FastText는 단어를 문자 단위 n-gram(예: 3-gram, 4-gram 등)으로 분해하여 임베딩을 학습합니다.
   - 예를 들어, "시나브로"라는 단어는 ["<시나", "시나브", "나브로", "브로>"]와 같은 subword로 쪼개집니다.
   - 이 방식 덕분에 **오타가 있거나, 학습 데이터에 없는 희귀 단어(Out-of-Vocabulary, OOV)**도 subword 조합을 통해 임베딩 벡터를 생성할 수 있습니다.
   - 예시: "apple"과 "applle"(오타)도 상당히 유사한 subword를 공유하므로, 임베딩 벡터가 비슷하게 나옴.
2. **형태소가 중요한 언어(한국어 등)에 유리**
   - 한국어처럼 어미, 접두사, 접미사 등 다양한 형태소 변화가 많은 언어에서 subword 기반 임베딩이 효과적입니다.
   - 예: "먹다", "먹었다", "먹고", "먹는" 등은 공통된 subword("먹")를 공유하므로 의미적으로 가까운 벡터를 가짐.
3. **학습 및 추론 속도가 빠름**
   - Word2Vec과 유사한 구조이지만, 효율적인 구현 덕분에 대용량 코퍼스에서도 빠르게 학습할 수 있습니다.
4. **텍스트 분류에도 활용 가능**
   - FastText는 단어 임베딩뿐 아니라, 문서 분류, 감정 분석 등 다양한 텍스트 분류 작업에도 효과적으로 사용됩니다.

FastText의 구조적 특징

- **입력**: 문장을 단어로 분리한 뒤, 각 단어를 다시 문자 n-gram(subword)으로 분해
- **임베딩**: 각 subword에 대해 임베딩 벡터를 학습
- **단어 벡터 생성**: 단어의 subword 임베딩 벡터를 모두 합산(또는 평균)하여 최종 단어 벡터로 사용

요약

- FastText는 **subword(부분단어) 정보를 활용**하여 오타, 신조어, 희귀 단어 등에도 강인한 임베딩을 제공합니다.
- 한국어처럼 형태소 변화가 많은 언어에 특히 효과적입니다.
- 빠른 학습 속도와 텍스트 분류 등 다양한 자연어처리 작업에 활용할 수 있는 실용적인 임베딩 기법입니다.