

■ 연구논문 요약문1

| | |
|---------------------------|--|
| <p>논문제목</p> | <p>Learning of indiscriminate distributions of document embeddings for domain adaptation</p> |
| <p>게재정보</p> | <p>Intelligent Data Analysis, vol. 23 no. 4, 2019</p> |
| <p>개요</p> | <ul style="list-style-type: none"> - 라벨이 정의되지 않은 새로운 문서 데이터에 적합한 머신러닝 모델을 구축할 때 기존에 가지고 있던 데이터를 활용하여 모델을 만드는 문제를 도메인 적응이라고 함 - 본 연구진은 텍스트 데이터에서의 도메인 적응 기술 개발을 위해, 도메인 적응에 적합하게 텍스트 데이터를 벡터화 시키는 단어 임베딩 기술을 개발함. 기존에 사용되면 BoW, TFIDF, DBOW, DM등의 방법론들은 텍스트 문서를 벡터화 시키는건 효과적으로 하였지만, 서로 다른 도메인들을 비슷한 공간에 임베딩 시키지 못하였음. 본 연구진은 Negative sampling을 할 때, 각 도메인에서의 노이즈와 비교하는 모델을 만드는 방식으로 새로운 단어 임베딩 기술을 개발함. |
| <p>연구결과</p> | <div style="text-align: center;"> </div> <p style="text-align: center;">도메인 적응 단어 임베딩의 시각화 그림</p> <ul style="list-style-type: none"> - 기존의 방법론을 사용하면 위의 그림과 같이 기존의 방법론들은 빨간색으로 나타나는 소스데이터와 파란색으로 나타나는 타겟데이터가 서로 분리되는 양상을 보이는데, 본 연구진이 개발한 방법론을 사용하면 오른쪽처럼 두 개의 데이터가 서로 섞이는 양상을 보임. - 이렇게 데이터가 섞이면 새로운 데이터에서 잘 적용이 될수 있는 도메인 적응 모델이 개발 가능함. |
| <p>활용분야 및 기대효과</p> | <ul style="list-style-type: none"> - 본 연구에서 개발된 모델을 활용하면 기존에 라벨이 없어서 기계학습을 적용하지 못했던 수 많은 문서 데이터들에 대하여 기계학습 모델을 구축 할 수 있음. - 따라서 이를 활용하면 문서 데이터에서 기계학습의 활용을 증대 시킬 수 있음. |