**한국어의 특성을 고려한 한국어 문서 분류 시스템 성능 개선**

2016131529

언어학과

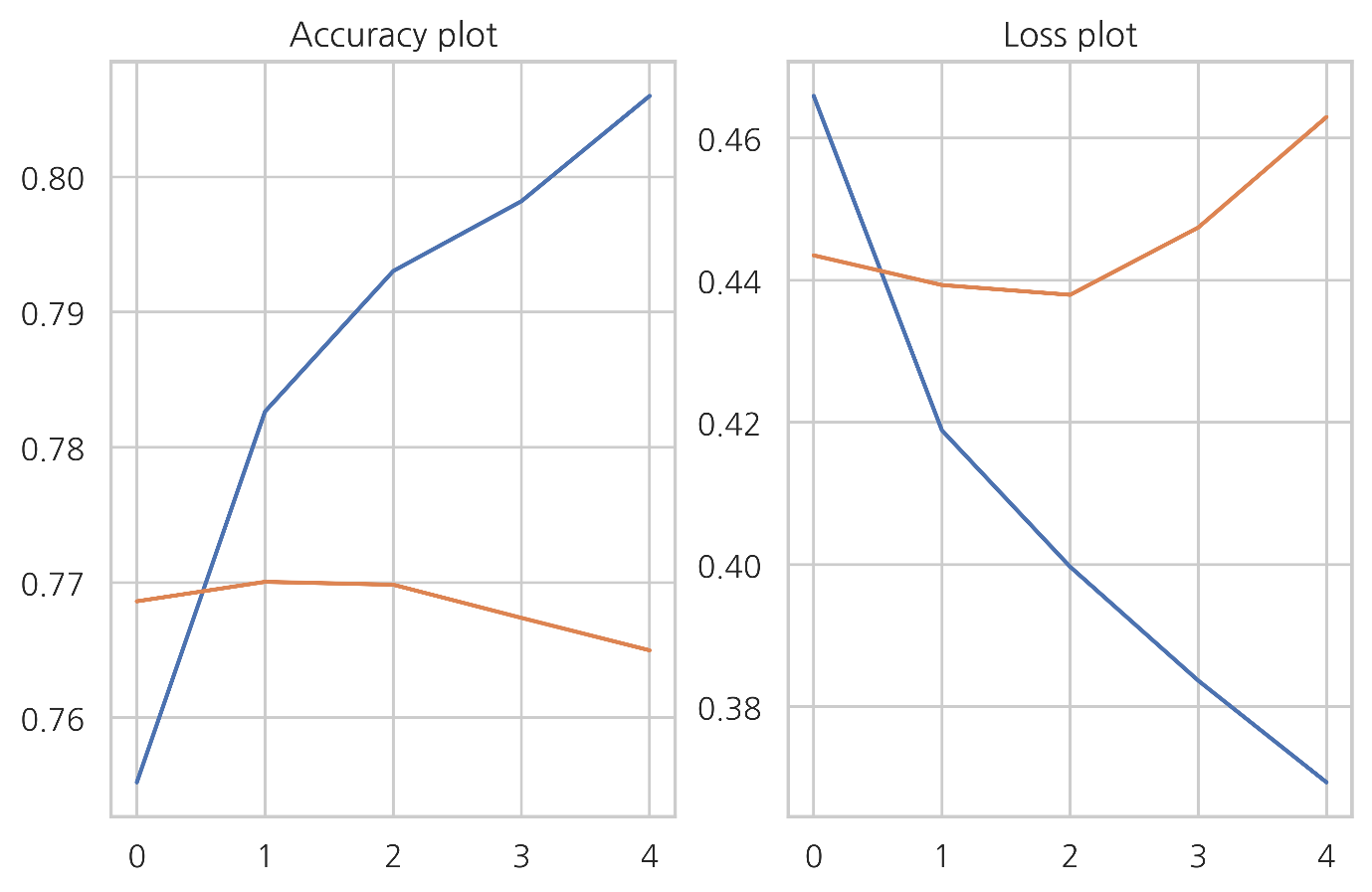
정찬영

**1. 서론**

문서 분류 시스템 시연 발표에서 저는 영어 스팸메일 데이터셋을 분류하는 작업을 해보았습니다. 영어의 경우에는 단어 수준의 토큰화만으로도 훌륭한 성능을 보이는 것을 확인했습니다. 그러나 한국어에서도 같은 알고리즘이 같은 성능을 발휘할 것이라고는 장담할 수 없습니다. 따라서 여기서는 우선 일전에 사용된 영어 스팸메일 분류기를 그대로 사용하여 한국어 영화 댓글 데이터셋에 분류해보도록 하겠습니다. 그 후 성능을 평가하여 문제점을 파악하고 소스코드를 개선하여 더 좋은 성능을 이끌어 내보도록 하겠습니다.

(보고서의 분량을 고려하여 활용한 소스코드를 첨부하였습니다.)

**2. 본론**

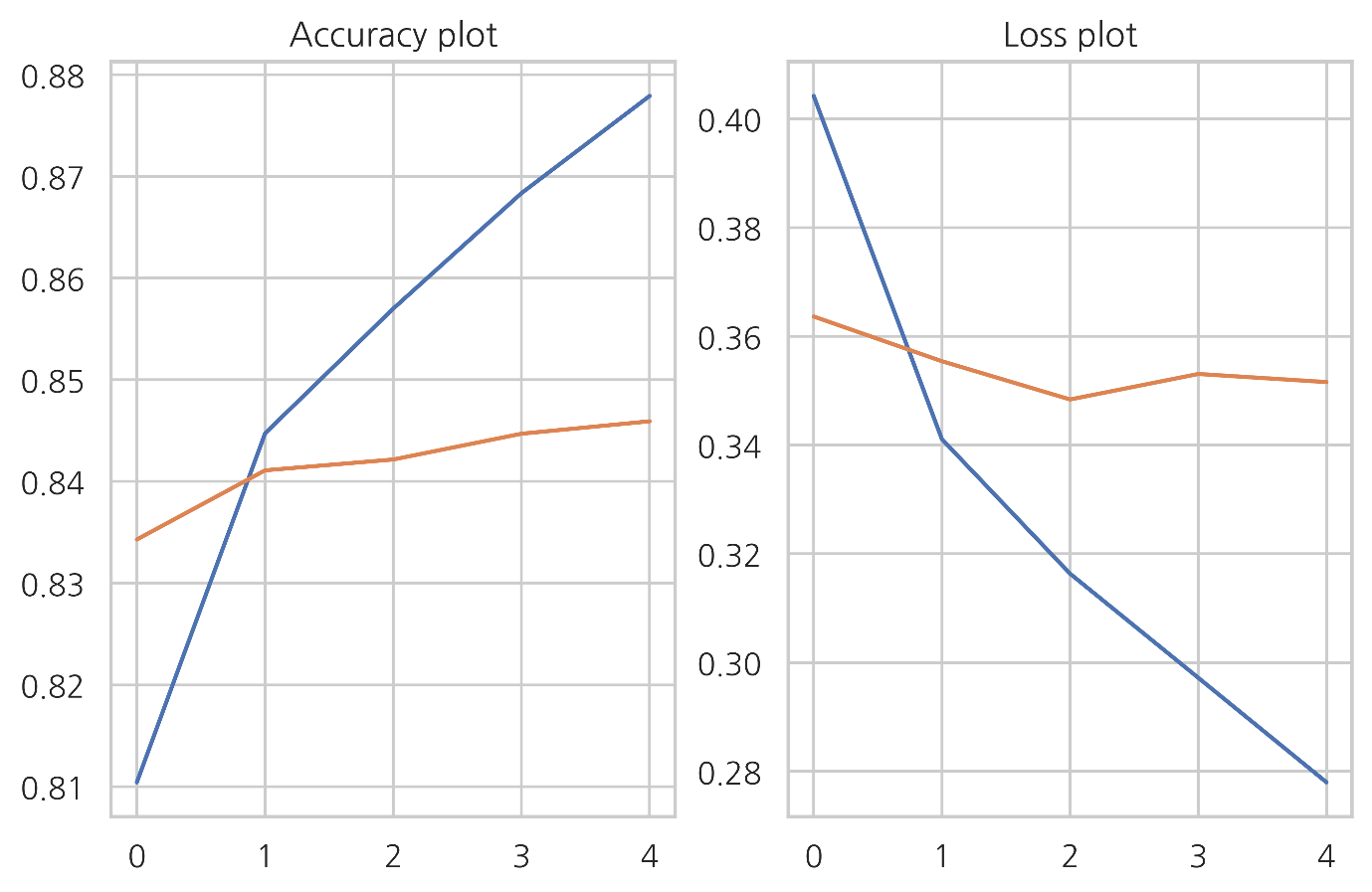
 영어 스팸메일 분류기를 이용한 결과는 다음과 같았습니다.

이미 약 1~2 에포크 이후 과대적합 양상을 보이며 정확도도 최대 77% 수준밖에 보이고 있지 않습니다. 이렇게 영어 텍스트에서 잘 작동했던 분류기가 한국어에서 잘 작동하지 않는 것은 희소성에서 기인하는 듯합니다. 영어는 고립어이기 때문에 어순에 의해 의미가 결정됩니다. 그러나 한국어는 교착어이기 때문에 어간과 접사가 붙어 의미를 형성합니다. 게다가 보조 용언, 조사 등 의미를 미세하게 변경하고 단어의 형태를 변화시키는 요소들이 많이 있기 때문에 단어 단위로 분석할 시 희소성이 커지게 됩니다.

즉 훈련세트에서 ‘먹었으니’라는 단어에 어떤 인덱스가 부여되고, ‘먹었어’에 다른 인덱스가 부여되었으나 테스트세트에서 ‘먹고’라는 단어가 나올 경우 훈련세트에서 전혀 보지 못한 단어로 판단하게 됩니다.

일전에 배운 형태소 분석기를 사용하는 것도 하나의 방법이겠지만 저는 여기서 서브워드 분절을 사용해보기로 했습니다. 서브워드 분절은 우선 모든 문장을 음절 단위로 분리한 뒤 확률이 높은 조합을 차례로 병합하여 토큰화를 수행합니다.

서브워드 분절을 수행하여 분류를 다시 한번 수행했을 시 성능은 다음과 같습니다.

 이전에 단어 단위로 분절을 수행했을 때 0.77 정도의 정확도를 보였던 것과 비교해보면 서브워드 분절을 수행했을 시 약 0.84 정도의 정확도를 보이며 전반적인 성능 향상이 이루어진 것을 알 수 있었습니다. 또한 5번의 에포크 이후에도 검증세트에 대한 정확도나 손실이 개선될 여지가 있어 보이기 때문에 파라미터를 조금 조정하여 과적합을 막아준다면 더 좋은 성능을 낼 수 있을지도 모릅니다.

**3. 결론**

여기서는 한국어와 영어의 차이에 입각하여 문서 분류기에 성능을 개선해 보았습니다만, 조금 더 생각을 확장해본다면 다른 언어별로도 접근법이 다를 것이고, 특히 지금까지 다루어 보지 않은 굴절어 같은 경우에는 더더욱 신중하게 접근해야 할 필요성이 있을 것입니다.

간단한 모델링이 아닌 최신의 정밀한 기술도 모든 언어를 아우르는 최적의 솔루션을 제공하지는 못한다는 사실은 나아가 인문학적 소양과 통찰력의 필요성을 반증하는 것이기도 합니다. 따라서 언어학적 관점에서의 문제 접근은 인공지능 시대에서 어떤 경쟁력을 제공하는가에 대해 지속적으로 성찰해볼 필요가 있습니다.