딥러닝을 활용한 뉴스 메타 태깅

콘퍼런스 발표 | 김기도 olaf.kido@kakaocorp.com

글 | 김규형 civtale@gmail.com 고려대학교 대학원에서 전기전자공학을 공부하고 있으며 딥러닝 기반 자연어 처리에 관심이 많습니다. 성향 분석이나 이슈 검출을 자연어와 연결시켜 의미 있는 결과를 만들어내는 것을 좋아하고, 데이터가 많은 곳을 직접 찾아다니려는 성향이 강합니다. 명확하지 않은 분야의 분석 모델을 구축하는 것을 좋아합니다.

메타데이터(Meta-data)

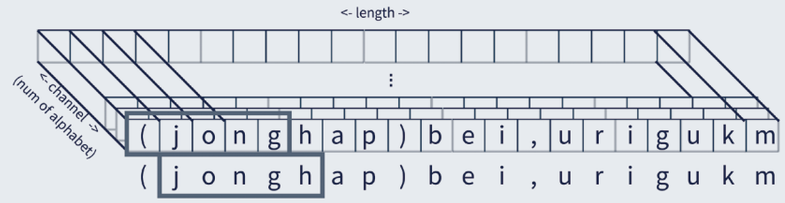
메타데이터는 뉴스 기사의 속성을 설명하기 위한 데이터로, 일반적인 정보들은 언론사가 카카오에 제공하기도 한다. 하지만 기사를 분석해야만 정의할 수 있는 정보들은 언론사에서 제공하지 않고, 제공할 수도 없는 경우가 있다. 카카오는 이런 경우에 해당되는 속성 태그를 생성하기 위해 직접 메타데이터를 만들어야 하는 번거로움을 겪는다. 사용자의 필요를 세밀하게 충족시키고 플랫폼에서 이를 활용하기 위해서는 뉴스의 메타데이터를 쉽게 파악할 필요가 있다. 카카오는 플랫폼을 통해 메타데이터를 소비할 수 있는 데이터의 흐름을 만들어 시스템화하고, 이를 '뉴스 메타 태깅 시스템'이라 이름 지었다.

기사의 형태 분류

카카오는 뉴스 기사를 서술된 내용을 기준으로 '사실 전달형 기사'와 '해설 묘사형 기사'로 나누어서 접근하였다. '사실 전달형 기사'는 "언제 누구에게 어떤 일이 있었다"를 육하원칙에 따라 서술한 기사이다. 반면 '해설 묘사형 기사'는 특정한 소재에 대해 맥락을 해설하거나, 현장을 상세하게 묘사하는 등의 심층 취재가 포함된 기사로 정의할 수 있다. 특히 '해설 묘사형 기사'는 '사실 전달형 기사'에 비해 상대적으로 포함된 정보가 다채롭거나 더 깊은 맥락을 제공하는 경우가 많아, 독자들이 깊이 읽어볼 만한 가치가 높은 특성을 가진다.

기사의 형태를 분류하기 위해 도입한 '합성곱 신경망(convolutional neural network, CNN)'은 6개 층의 '컨볼루셔널 레이어(convolutional layer)'와 3개 층의 '완전 연결 레이어(fully connected layer)'로 구성되어 있다. 텍스트의 형태소를 분석하거나 워드 임베딩(word embedding)을 하지 않고 그대로 분석할 수 있는 형태이면서 기사 안에서 사용되고 있는 괄호나 인용부호 등의 기호까지 의미 있게 반영하기 위하여 글자 단위로 학습을 진행하게 되었다. 또한 글자 단위에서 차원이 작은 상태로 처리하기 위해 한글과 한자를 모두 로마자 표기법(Romanization)으로 변환하여 학습 자료로 구성하였다.

[그림 1] 뉴스 기사 콘텐츠 분석을 위한 고려 사항들



· 형태소 분석, Word Embedding보다는 캐릭터(Character) 단위 처리
  ▶ 기사 제목이나 본문의 특징적 기호 활용(예: [날씨], (종합), "나혼자")
· 한글, 한자 Romanization(로마자표기법 변환, 음역)
  ▶ alphabet space 안에서 처리(예: 文정부, 靑 "소득성장가속")
· 문자와 단어의 순서 보존하여 기사의 문투 학습 ==> 1D Convolution

카카오는 뉴스 기사의 형태 분류 정보를 갖고 있지 않다. 딥러닝(deep learning)으로 학습을 하기 위해서는 정답 세트(set)가 필요하지만, 정답을 가지고 있지 않기 때문에 정답 세트를 어떻게 만들 것인지와 이에 대한 답을 찾아야 하는 문제가 발생한다. 카카오는 이러한 문제를 해결하기 위해 뉴스 기사를 파악한 결과값이 500자 내지 1200자면 사실 전달형 기사로 정의하고 3000자 이상의 기사는 해설 묘사형 기사로 정의하며, 뉴스 기사의 길이로 단위를 나누어 속성을 분류하였다. 잘못된 분류가 있다고 하더라도 오분류된 자료들에 특정한 패턴이 없다면 큰 단위에서 전체적인 학습을 하는 대세에는 영향을 주지 않는다. 따라서 자료의 양이 많은 트레이닝 세트(training set)는 자동으로 정답을 구성하고, 테스트 세트(test set)에 대해서는 전문가들이 직접 감수한 정답 세트로 검증하고 있다.

테스트에서 입력 차원의 크기는 중요한 정보들이 기사의 앞에 올 것이라는 가정을 세워서 제목과 본문을 합치고 로마자화하여 앞 600자로 구성하였다. 단어 차원의 크기는 알파벳 및 문장부호를 원핫 인코딩(one-hot encoding)으로 처리하여 그 크기를 정의하였다.

먼저 첫 번째 테스트는 사실, 해설, 기타 데이터 세트의 비율을 10:1:10으로 진행하였다. 정확도(accuracy)는 약 80%에 육박했지만, 해설형 기사의 데이터가 부족했기 때문에 해설 카테고리의 재현율(recall)은 43% 정도에 그쳤다. 데이터를 추가하고 2:1:2 비율의 데이터 세트를 구성하여 두 번째 테스트를 진행한 결과 정확도 82%, 재현율 62%의 수준을 보였다. 더 나아가 동일한 기사를 1014자를 사용하여 학습한 결과, 정확도가 81% 정도로 기존 대비 1% 떨어졌으나 해설 묘사형의 재현율은 75%로 상당히 향상되었다. 한편 연재나 기획 쪽 기사를 판별하는 능력은 향상되었으나, 오늘의 운세와 같은 종합 기사에서는 4000자가 넘어가는 경우도 있었기 때문에 학습 및 판별에서 혼동하는 양상이 발견되었다.

앙상블 모델(ensemble model)은 여러 개의 모델 결과를 조합해 더 좋은 결과를 구성하도록 하는 모델 구조를 말한다. 위의 테스트는 두 모델의 판정 결과로 나온 소프트맥스(softmax) 확률값을 더하여 카테고리를 분류하는 구조로 구성하였다. 전문가들은 위의 테스트 결과에 대해 상당히 좋은 평가를 내렸지만, 이와 비교할 대상이 없다는 것이 다소 아쉬움을 남긴다.

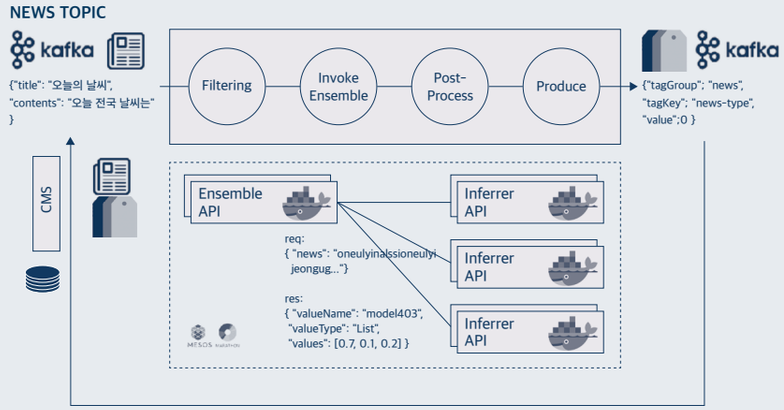
학습에서 원만한 분류가 이루어지지 못하는 예로 '외국어를 공부하는 코너 기사'를 들 수 있다. 해당 예에서는 외국어 문자가 삽입되어 있는 기사들을 적절히 판단할 수 있는 데이터가 부족하여 정확한 판정을 내릴 수 없는 문제가 나타난다. 이에 해당 데이터 자료에 대한 학습을 진행하여 결과물을 개선할 수 있는 방법이 필요하다.

뉴스 메타 태깅 시스템

카카오에서는 태깅 시스템을 개발하면서 세 가지 설계 목표를 제시하였다. 첫째, 기사가 들어오는 대로 메타데이터를 빠르게 생성해야 하는 특성상 실시간(real-time)으로 작업이 진행되어야 한다. 둘째, 작업의 의존관계(dependency)가 없는 경우 여러 작업자들이 동시에 시스템에 접근해 메타데이터를 생성할 수 있도록 해야 한다. 이것을 비동기적 작업이라고 하며, 특정한 작업에서 문제가 발생했을 경우에도 나머지 작업 전체가 영향을 받지 않도록 하여 안정성을 높일 수 있다.

마지막으로 어떤 알고리즘이나 프레임워크(framework)를 사용하더라도 시스템이 작동하도록 설계해야 한다. 뉴스 본문을 분석하기 위해 사용하는 기계학습 및 딥러닝 기법에 제약을 두지 않기 위해서다.

[그림 2] 기사 형태 분류 시스템



카카오 뉴스의 태그 처리 시스템은 [그림 2]와 같은 절차를 거쳐 관리된다. 뉴스 데이터를 관리하는 카카오 CMS 플랫폼에서는 실시간으로 들어오는 뉴스를 뉴스 토픽(news topic)으로 전송한다. 뉴스 토픽에서 기사를 꺼내어 메타 태그를 생성할지 판단하고 태깅 서비스를 호출하는 일은 데이터 흐름 제어부가 담당한다. 생성하는 메타데이터의 종류가 많아질 수록 호출해야 할 태깅 서비스가 늘어나고 데이터 흐름 제어의 역할이 반복되므로 이 부분은 아파치 나이파이(Apache NiFi)를 사용하여 구현하였다. GUI 기반 환경에서 데이터 흐름을 구현할 수 있기 때문에 개발 시간을 단축할 수 있고, 데이터의 흐름을 메시지 단위로 추적하기 용이하기 때문이다.

태깅 서비스는 앙상블(ensemble) API와 인퍼러(inferrer) API로 이루어져 있다. 앙상블 API는 인퍼러 API들의 결과 값을 종합하여 뉴스 기사에 대한 메타 태그값을 결정한다. 앞서 설명한 기사의 형태 분류의 예라면 기사가 사실 전달형 기사인지 해설 묘사형 기사인지가 메타 태그의 값이 된다. 이렇게 결정된 메타 태그는 태그 토픽(tags topic)으로 전송한다. 여기서 메타 태깅 과정을 끝날 것 같지만 한가지 과정이 더 남아있다. 카카오 CMS 플랫폼은 태그 토픽에 저장된 메타 태그와 뉴스 본문 데이터를 병합하여 다시 뉴스 토픽으로 발행시킨다. 이와 같이 태그 데이터의 흐름을 구성하게 되면, 뉴스 토픽을 소비(consume)하는 것만으로도 뉴스 콘텐츠와 부가 생성된 메타 태그를 한 번에 받아볼 수 있게 되어 소비자(consumer)들이 데이터를 여러 번 조회하지 않게 할 수 있다.

태깅 서비스를 구현하기 위해서 스파크 스트리밍(Spark Streaming)이나 스톰(Storm)같은 실시간 데이터 처리 프레임워크를 검토하기도 하였지만 특정 기술 세트에 대한 종속을 피하기 위해 필요한 역할에 따라 각기 독립적인 웹 API를 개발하였다. 개발된 API는 앙상블 API와 인퍼러 API로 구분되며, 그 설명은 다음과 같다.

앙상블 API(ensemble API)

앙상블 API는 말 그대로 학습한 모델들이 내어놓는 추론 값들을 종합하여 최종 결과 값을 반환하는 API이다. 앞서 설명한 기사의 형태 분류의 예라면 뉴스 기사를 로마자로 변환한 뒤, 600자 기준 판별 모델의 추론 값과 1014자 기준 판별 모델의 추론 값을 합하여 최종 결과 값을 내게 된다. 또한 인퍼러 API의 게이트웨이(gateway)로서 인증이나 로깅과 같은 공통 역할을 수행하기도 한다.

인퍼러 API(inferrer API)

인퍼러 API는 학습한 모델을 사용하여 데이터 추론을 하는 API로, 모델의 개수만큼 API를 개발한다. 앞서 설명한 기사의 형태 분류에서는 600자 기준 판별 모델과 1014자 기준 판별 모델이 각각의 인퍼러 API이다. 인퍼러 API는 최대한 단순하게 구현하는 것을 지향하여 다른 기능 없이 오직 모델의 추론 기능만을 구현한다. 모델이 재학습 되는 경우가 있으므로 버전 관리가 중요한데, 학습한 모델과 코드를 통합하여 관리하기 위해 도커라이징(dockerizing)을 하여 배포하고 버전을 관리한다.

뉴스에는 텍스트뿐 아니라 이미지, 차트, 동영상 등이 포함되어 있기 때문에, 향후 카카오에서는 텍스트뿐 아니라 이러한 멀티미디어 자료들을 분석할 수 있는 시스템을 개발할 예정이라고 한다.

〈딥러닝을 활용한 뉴스 메타 태깅〉 브런치로 연결되는 QR 코드입니다.