MBC 장애인용 폐쇄자막 아카이브 시스템

글. 노현우 MBC 방송IT센터 유통인프라부

2020년 2월 장애인용 폐쇄자막 아카이브 시스템(이하 자막시스템) 개발이 완료되었다. 유통인프라부는 송출신호와 연동하여 송출한 콘텐츠와 동기화된 자막을 생산, 생산한 자막은 사내 부서에 송신, 자막 메타데이터를 활용한 검색, 자막과 영상 간의 시간 조절 및 편집 기능을 개발하여 자막시스템을 운영 중이다.

개발 배경

기존 자막시스템은 2012년에 구축되어 운영에 어려움이 있었다. 응용프로그램 소스코드의 유실로 유지보수가 불가하여 기능개선이 어려웠다. 기존 자막시스템은 여의도에서 사용하던 송출시스템과 연동되어 PCM 프로그램에 대한 대응이 되지 않아 영상과 자막의 타임코드가 일치하지 않았다. 그리고 CS(Client/Server) 프로그램으로 개발되어 유지보수 및 배포가 어렵고, 특정 OS에 종속적인 문제가 있었다. 그리고 자막의 보정 기능이 필요하였다. 자막은 속기사가 영상의 음성을 듣고 실시간으로 작성하여 영상과 자막 간의 미세한 시간 차이가 발생하거나, 오타가 간혹 발생하기 때문에 보정 기능은 필수적이다.

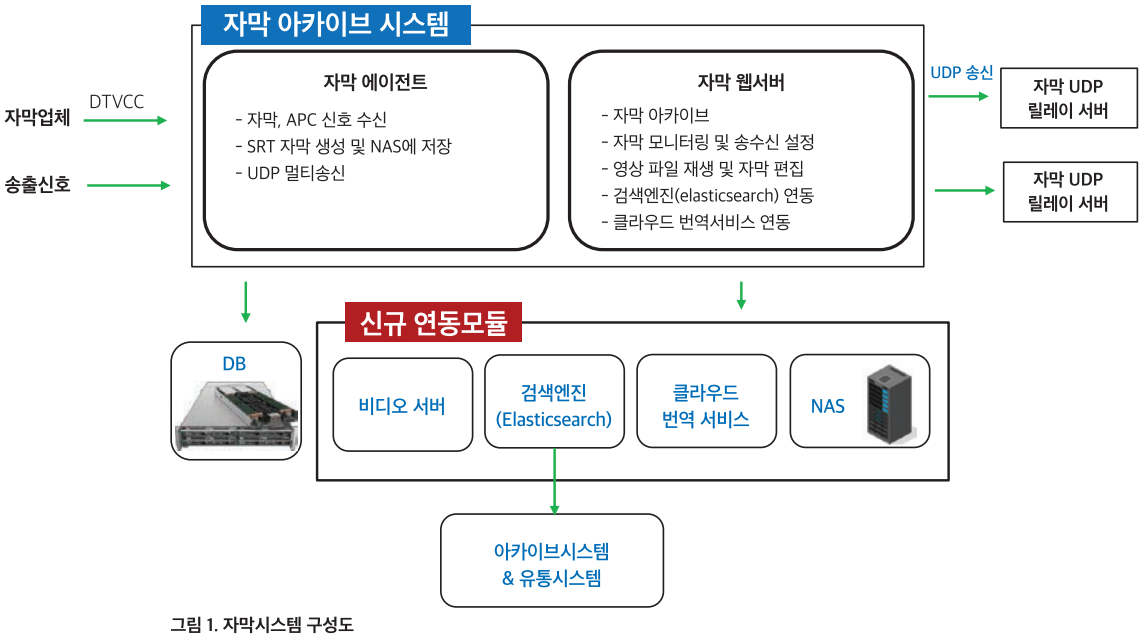

그림 1. 자막시스템 구성도

아카이브한 자막은 양이 방대하여 사용 중인 검색엔진에 색인할 수 없었고, 색인하는 경우 검색엔진에 부하를 발생시켜 검색이 정상적인 응답속도로 실행되지 않았다. 이러한 문제를 해결하기 위한 방안이 모색되었고, 자막만을 위한 검색엔진을 구축하기로 하였다. 그리고 콘텐츠제작부서에서 다국어 자막에 대한 요구사항이 있어 한글자막을 활용한 영어자막 생성이 필요했다. 클라우드 번역 서비스를 활용한 영어번역 서비스는 완벽하진 않지만, 번역 초벌용으로 의미가 있다고 판단하여 개발하기로 하였다.

자막시스템 개발

자막시스템은 자막에이전트와 자막웹서버로 구성되어 있다. 자막에이전트는 자막제공업체로부터 자막을 수신하고, 송출신호를 수신하여 영상과 타임코드가 동일한 자막을 생성하는 기능, 영상모드(완제, 합본, 분리본, 24시간)별 자막생성 기능, 자막 멀티 송신기능 등이 개발되었다. 아카이브 모드를 기준으로 프로그램은 일반프로그램, PCM 예능프로그램, PCM 드라마프로그램으로 나누어질 수 있다. 일반프로그램은 완제 모드로, PCM 예능프로그램은 합본모드, PCM 드라마프로그램은 완제 모드로 아카이브 하는데 이러한 모드에 대응되도록 자막 생성 기능을 구현하였다. 그리고 장애에 대처하기 위해 24시간 자막생성 기능도 같이 개발하였다. 기존 시스템은 하나의 CS 프로그램에서 한 군데만 송신할 수 있어 제약이 많았었다. 그래서 수신한 자막을 한 번에 여러 군데로 송신하는 멀티 송신 기능을 구현하였다.

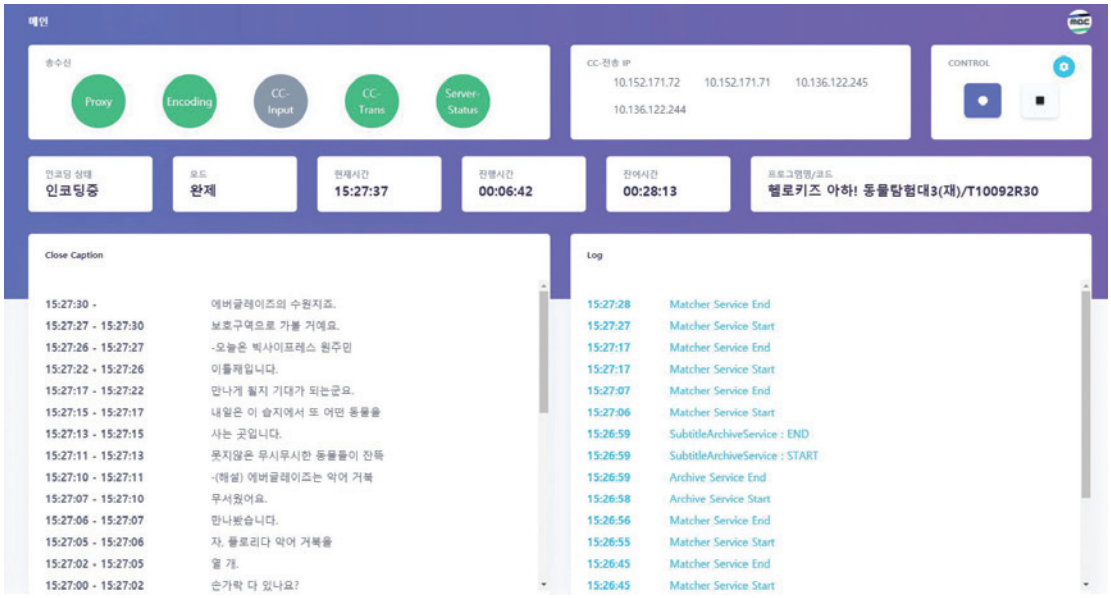

그림 2. 자막에이전트가 수신한 자막과 송출신호 모니터링 페이지

자막웹서버는 자막에이전트로부터 자막을 수신하여 웹에 표출하는 기능, 자막에이전트 모니터링 기능, 자막 수신 모니터링 기능, 송출 신호 모니터링 기능, 자막 인코딩 모드/자막 저장 경로/전송처 설정 기능, 자막 수동 등록기능, 자막-영상 매핑기능, 자막 보정 기능, 자막 다운로드 기능, 자막 아카이브 기능, 검색엔진 연동 기능, 클라우드 번역 서비스 연동 기능을 개발하였다. 기존 자막 모니터링은 CS 프로그램으로 이루어져 단말마다 CS 프로그램을 설치해야만 모니터링 가능했으나, 웹으로 구현하여 모든 단말에서 모니터링 가능하다. 자막은 영상의 음성을 듣고 생성되기 때문에 타임코드가 미세하게 영상과 불일치할 수밖에 없다. 또한 자막은 오타가 발생할 수도 있다. 이러한 문제점 때문에 자막 보정이 필요한 경우가 종종 발생하는데, 이런 경우 영상을 보면서 자막을 보정할 수 있는 기능을 개발하였다. 영상은 아카이브 시스템과 유통시스템에 사용하는 영상을 연동하여 자막과 함께 보여주어 자막 보정을 손쉽게 할 수 있도록 구현하였다.

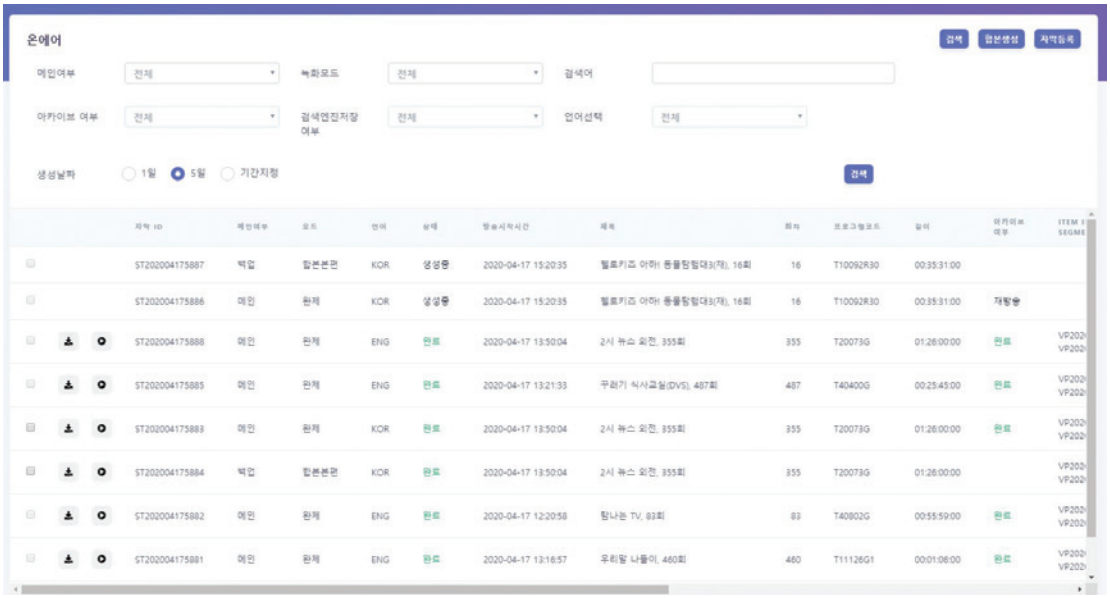

그림 3. 자막시스템이 생성한 자막을 관리하는 페이지. 자막의 프로그램, 회차, 생성모드, 언어, 생성 여부 검색 및 표시

사내 제작 부서에서 다국어 자막에 대한 요구사항이 있어 클라우드 번역 서비스를 활용한 영어자막을 생성했다. 영어자막은 사업용으로 바로 사용하기는 부족하나, 초벌본 영어자막을 생성했다는데 의미를 두고 점점 정확도를 높여갈 예정이다. 아카이브한 자막은 아카이브시스템과 유통시스템에서 검색할 수 있도록 자막시스템 전용 검색엔진에 저장하는 기능을 개발하였다.

그림 4. 영상 프리뷰 페이지. 자막 수정, 자막 타임코드 수정, 영상과 자막의 매핑

자막시스템 전용 검색엔진(Elasticsearch) 구축

아카이브시스템과 유통시스템은 상용 검색엔진을 사용 중이다. 검색은 점점 고도화되어 대부분의 메타데이터를 활용하여 검색이 되고 있으나, 자막은 데이터가 방대하여 검색엔진에 색인하면, 검색엔진에 부하를 주어 응답속도가 현저히 떨어졌다. 자막 검색만을 위해 상용 솔루션을 추가 구축하는 것은 무리가 있어 오픈소스 검색엔진

인 Elasticsearch를 구축하기로 했다. Elasticsearch는 'nori'라는 한글 형태소 분석기를 지원하고 있어 한글을 검색하는데 무리 없을 것이라고 판단하여 선정하게 되었다.

검색 시스템의 구성은 수집기, 스토리지, 색인기, 검색기로 구성되어 있다. 수집기는 단어를 수집하는 프로그램이다. 자막시스템이 생산한 자막을 Elasticsearch의 수집기에서 수집하고 수집한 자막은 색인기를 통해 형태소 분석 이후 색인한 데이터를 Elasticsearch 저장소에 저장한다. 저장한 색인데이터는 아카이브시스템과 유통시스템에서 검색을 요청하면 검색기가 저장소를 쿼리하여 검색 결과를 응답한다.

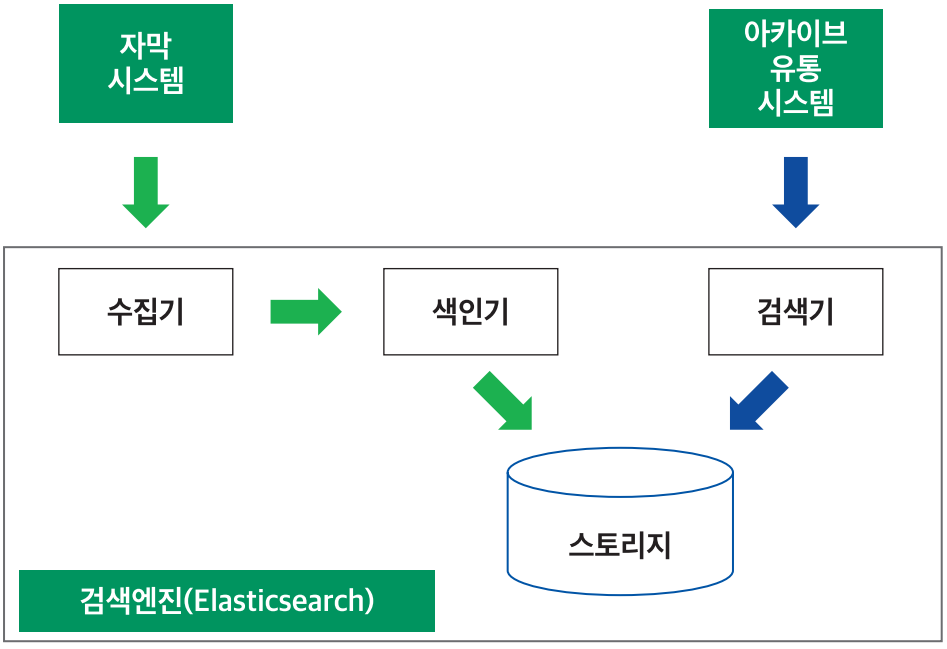

그림 5. 검색엔진의 구성도. 자막시스템이 검색엔진에 저장 요청을 하면 수집기를 통해 수집하고, 색인기에서 형태소 분석 후 스토리지에 색인되어 저장, 아카이브유통시스템에서 검색을 요청하면 검색기가 스토리지에서 단어를 검색하여 응답한다.

Elasticsearch의 저장소는 RDBMS와 다르게 다음과 같은 특징이 있다. 예를 들어, RDBMS에서 사용하는 Database는 Elasticsearch에서는 Index라는 용어를 사용하고, RDBMS는 Row를 Elasticsearch는 Document라는 용어를 사용한다. 영상의 자막은 하나의 Document로 Elasticsearch에 저장된다. Elasticsearch는 nested 타입을 지원하는데, 하나의 필드를 Object 객체 배열을 독립적으로 사용할 수 있다. 기존 RDBMS에서는 두 개의 테이블을 생성하고 외래

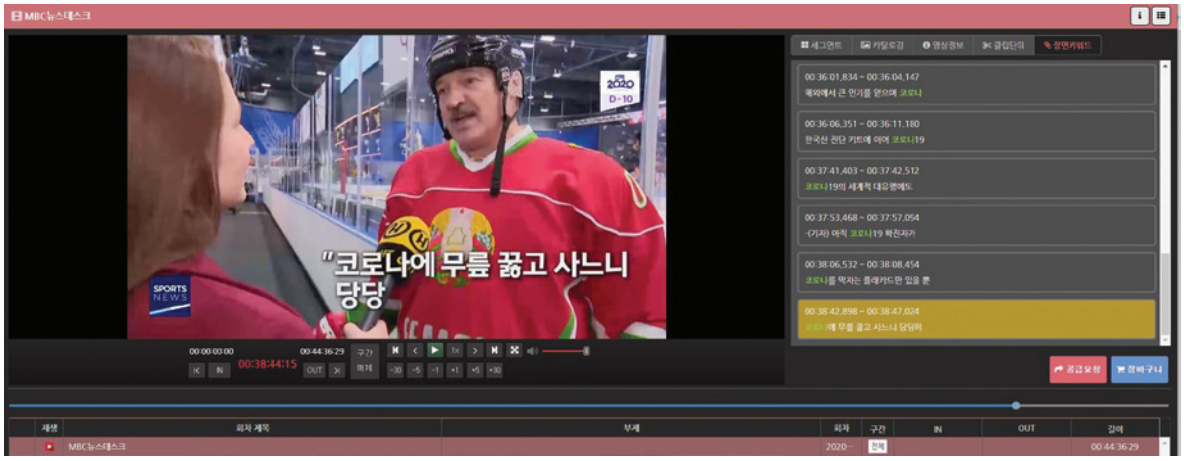

그림 6. 유통시스템 영상 프리뷰 페이지. 유통시스템과 연동하여 자막검색 기능 제공

키를 통해 관계를 맺었다면, Elasticsearch는 Document에 하나의 필드에 Object 객체 배열을 지원하여 한 번에 많은 데이터를 조회할 수 있다.
자막시스템과 연동하여 데이터를 저장할 때 자막제목, 영상 id, 자막 언어코드, 자막 object 객체 타입으로 설계하였다. 자막 object는 자막번호, 자막시작시간, 자막종료시간, 자막텍스트로 구성하였다. 그래서 검색어를 쿼리하면 검색어가 포함된 자막시작시간, 자막종료시간, 자막텍스트가 응답한다.

향후 계획

이번 시스템을 개발하면서 향후에 보완할 점이 몇 가지가 있다.
첫 번째는 한글 맞춤법 검사기를 구축하지 못했다. 한글 맞춤법은 API로 제공되지 않아 자막 맞춤법을 자동으로 검사할 수 없다는 점이 아쉬웠다. 향후 맞춤법 검사 API가 제공한다면, 자막 생성 시 맞춤법 기능도 구축할 예정이다.

두 번째는 영어자막이다. 영어자막에 대한 요구는 많으나, 정확성이 아직 많이 부족하여 사업적 사용 가치는 낮은 상태이다. 영어자막을 좀 더 개선할 수 있는 방안을 마련할 예정이다.

그리고 마지막 세 번째는 오픈소스 검색엔진(Elasticsearch) 운영이다. 오픈소스는 처음 구축 시에는 사용에 문제가 없지만 장애 발생 시 또는 세밀한 설정 시 숙련된 기술이 필요하다. 그래서 오픈소스를 도입하더라도 전문 유지보수업체에 의뢰하는 경우가 많다. 현실적으로 유지보수업체 의뢰하는 것은 힘들다고 판단되고, 기술 향상을 위한 학습이 필요하다고 생각된다. 그리고 기술이 갖춰지면 자막 검색뿐만 아니라, 사내 CMS 검색에도 적용하도록 하겠다.