1) 안녕하십니까
국어사전 DB를 사용한 Labeling tool 에 대한 발표를 하게 된 조우진입니다.

2) 먼저 목차를 보면 개발의도, 개발계획, 성능비교 순으로 진행하겠습니다.

3) 이러한 레이블링 툴을 제작하여야겠다고 생각한 이유는 첫째, 한국어 관련 레이블링 툴을 조사 당시엔 발견하지 못하였고,
둘째, 개체명 태깅과 관련하여 자동화를 해야겠다고 생각했습니다.
마지막으로, 수작업을 통하면 인력소모가 심각하여 원하는 양의 데이터를 확보하기 어렵다고 생각하였습니다.

4) 툴을 개발하기 위해서는 먼저 각종 데이터들이 어떻게 구성되어 있는지를 파악하는 것이 중요했습니다.
그래서 가장 먼저 파악한 점은 ETRI, 한국전자통신 연구원, 이하 에트리라고 하겠습니다.
에트리의 개체명 분류 기준은 다음과 같이 15개의 대분류와 160개의 세분류로 나누어집니다.

5) 다음으로 알아볼 것은 에트리의 형태소 분석, 개체명 인식 예시입니다. 다음과 같이 한 문장을 형태소 분석한 뒤, 개체명을 태깅하는 방식으로 진행됩니다.

6) 그래서 다른 학습데이터의 말뭉치 양식을 비교하였습니다. 국립국어원은 형태소별로 아이디별로 정렬한 뒤, 형태소 태그와 위치, 가중치 순으로 정렬되는 것을 볼 수 있습니다. 엑소브레인 말뭉치는 원문에서 개체명인 부분은 꺽쇠를 사용하여 태그를 해놓은 자료입니다. 작년팀이 사용한 데이터와 동일한 내용임을 확인하였습니다.

7) 울산대학교 말뭉치는 원문, 형태소 분석, 개체명 태깅 순으로 정렬되어 있습니다.

8) AI HUB 말뭉치는 일반 상식에 대한 말뭉치를 확인하여 학습용 데이터로 사용하기에는 부적합한것으로 확인되었습니다.

9) 다음으로 실제 개발에 이용할 표준국어대사전의 전문 분야입니다. 원래 추가하기로 했던 태그는 음식, 국가, 직업 이었지만 해당 태그를 학습데이터에 적용하기에는 많은 시간과 노력이 필요할거 같아 표준국어대사전의 분야 중 동물, 식물, ㅇㅇㅇ 를 늘릴 것으로 예상됩니다.
전년도 학습 데이터는 다음과 같은 양식으로 되어 있어, 이와 같은 양식으로 데이터를 가공할 것으로 예상됩니다.

10) 실제 개발 계획에 관해서 설명드리겠습니다. 왼쪽 사진은 실제 표준국어대사전의 데이터입니다. 저희는 해당 표준국어대사전의 DB 파일을 확보하였고, 이를 중간 사진처럼 단어와 분야로 텍스트화 할 것입니다.
오른쪽 사진은 저희가 실제 사용하게 될 개체명이 미태깅 된 텍스트 데이터입니다.

11) 표준국어대사전 DB 파일을 strtok 함수를 통해 공백을 기준으로 두 단어로 자릅니다. 그 후에, 표준국어대사전 DB의 단어 부분과 미태깅 된 텍스트를 strcmp를 사용하여 비교합니다. 같은 단어를 찾게 되면, 표준국어대사전 DB의 전문 분야를 strcat를 사용하여 미태깅 된 텍스트에 추가합니다.
이렇게 단어와 전문분야가 매칭된 텍스트는 태그를 영어로 수정하여 오른쪽처럼 가공하면 학습용 데이터로 쓸 수 있을 것으로 예상됩니다.

12) 마지막으로, 저희가 생각한 아이디어와 유사한 UTagger-NE라는 프로그램이 존재합니다. 개체명을 태깅할 수 있는, 즉 레이블링이 가능한 툴입니다. 이러한 툴을 사용하여 개체명을 추가, 삭제할 수 있고, 형태소 분석된 단어를 클릭 및 개체명을 수동으로 추가할 수 있습니다.
이러한 툴의 문제점은 데이터 파일 내의 단어 수가 현저히 부족하여 고유명사와 같은 부분에 대해서는 형태소분석 단계부터 제대로 되지 않는 문제점이 있었습니다.
그러한 문제점으로 인해, 최종적으로 해당 프로그램을 사용하기에는 부족합니다. 만약 설명드린 툴을 제작하게 된다면 이런 미흡한 부분을 개선하기 위하여 우리말샘 사전, 위키피디아 인명사전 등을 참고하여 제작할 것입니다.

아이디어를 발표하는 단계여서 많이 미흡한 점 양해 부탁드립니다.