



# Gcon - 신규기능 중간 이후 보고서

작성일: 2021년 1월25일

작성자: 오지혜


프로젝트 완성도 : **70%**/100%


## 목차

1. 데이터 추출 항목
2. 추출 방법
3. 결과
4. 개선 방향

---

## 1. 데이터 추출 항목

 파일명: **input.xlsx (예시)**

 추출 해야하는 데이터

**I am Seonho.**

**I am Sunho,**

**I am Seoho.**

첨부파일

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/fb7fb554-a2b4-44fc-983a-ca539fb7d647/input.xlsx>

- 위 그림과 같이, 문장 내에서 **다른게 표시된 단어에 태그 표시를 해서 추출합니다**

## 2. 추출 방법

### a. 3개씩 끊어 가져와서 리스트에 담기

`[[case_a(3문장)], [case_b] ..... ]`

### b. case 별로 각 문장을 형태소 분석을 한다.

"list > tuple > str" 형식으로 형태소 분석

#### 1) stanford 분석

소요시간이 190배 이상 걸리며 클래스도 너무 다양하여, 이미 잘 작업되고 있는 nltk로 작업했습니다.

#### 2) nltk try

I. 영 단어의 품사별로 나눠줬습니다

- '동사 + 전치사' 함께 태그 표시  
형태소 분석의 조합을 통해 이를 해결했습니다.
- example : **learn of = know**

II. **Mr. vs Mr, & don't & do not**  
비교하면 . 와 , 가 균등하게 분리되지 않는다. 따라서, 이러한 예러도 해결했습니다.

### 😊 Good Case

#### • Input Data :

How did you **learn of** this position?

How did you **know** this position?

How did you **get** this position?

#### • Output Data :

How did you **<b>learn of</b>** this position?

How did you **<b>now</b>** this position?

How did you **<b>get</b>** this position?

## 3. 결과

Case	Input	Output
1	I am Seonho.	I am <b>seonho</b><b>.</b>
	I am Sunho,	I am <b>sunho</b><b>,</b>
	I am Seohoo.	I am <b>seohoo</b><b>.</b>
2	I need to drink something	I need to <b>drink</b> something
	I need to drunk something	I need to <b>drunk</b> something
	I need to drank something	I need to <b>drank</b> something
3	I like this person.	I like <b>this</b> person.
	I like that person.	I like <b>that</b> person.
	I like those person.	I like <b>those</b> person.
4	How did you learn of this position	How did you <b>learn of</b> this position
	How did you know this position	How did you <b>know</b> this position at
	How did you get this position	How did you <b>get</b> this position at

a. case set 수 : 29쌍 (87줄) < - 부족하므로 데이터를 가져올 예정입니다.

b. 소요시간: 1초

c. 첨부파일

[https://s3-us-west-2.amazonaws.com/secure.notion-static.com/8c778eb3-28b9-4e08-a156-4bbf251aab6b/input\\_result\\_01251036.xlsx](https://s3-us-west-2.amazonaws.com/secure.notion-static.com/8c778eb3-28b9-4e08-a156-4bbf251aab6b/input_result_01251036.xlsx)

## 4. 개선방향

### a. 데이터 샘플 수집 및 정확도

데이터의 개수가 너무 적기 때문에 한국어 관용 표현을 구글과 파파고에서 번역 후 다르게 번역된 부분을 찾아서 표시 해주는 등의 작업을 통해 무작위의 데이터 속에서도 스크립트가 정확하게 작동하는 개선 작업이 필요합니다.

[시간이 조금더 소요된 가장 큰 원인은 같은 형태, 품사, 사용법일지라도 형태소 분석을 하면 각기 다르게 인식되는 경우가 예외로 있었습니다. (e.x. [of, learning] → [전치사, 명사] // [of, going] → [전치사, 동사])]

따라서, 이러한 예외를 잡기위해 더 많은 결과물을 본 후 진행 할 수 있다고 생각했습니다.

## **b. 개선 사항**

서지혜 매니저님과 잠시 짧은 소통을 한 후, 더 많은 무작위의 데이터를 가져오기 위해, **한국어 관용표현 등을 구글, 파파고 로 번역하고 이 결과값을 비교해서 다른 부분을 표시**해주는 작업에서 테스트를 진행하고 스크립트를 더 정확히 작업을 하고자 합니다!

앞으로는 단순히 표시되는 단어만 다른 부분을 찾지 않고, **같은 의미 이겠지만 다르게 번역된(표현된) 각 문장마다 사용되는 특정 부사표현과 명사 단어를 함께 가져와서 <b></b>로 표기되는 부분을 구글에서 표현검색("") 진행했을 때도 매끄럽게 결과가 나올 수 있도록** 하면 어떨까 하는 아이디어를 생각해 보기도 했습니다!