텍스트 기반의 훈련 데이터 구축을 위한 자동 데이터 태깅 작업에 대한 연구

김나연*, 소혜령*, 박준호*
*(주)솔루게이트 기업부설연구소
e-mail: {na2na8, so1716, park1058}@solugate.com

A Study on Automatic Data Tagging for Text-based Training Data Construction

NaYun Kim*, Hyeryung So*, Joonho Park*
*Research and Development Laboratory, Solugate Ltd.

요 약

텍스트 기반의 훈련 데이터는 데이터를 수집한 이후에 각 문자별로 태깅 작업이 필요하다. 말뭉치(Corpus)는 언어학에서 주로 이루고 있는 텍스트 집합이다. 말뭉치는 각 단어의 품사 표기에 대한 정보가 태그 형태로 되어 있다. 본 연구에서는 한국어 기반의 태깅 작업을 연구했으며, 기본 한국어 말뭉치가 아닌 기업이나 연구 기관에서 데이터를 수집하여 말뭉치나 별도 학습 데이터를 구축하기 위한 자동 태깅 방법에 대해 알아본다.

1. 서론

최근에는 한국어 데이터 수집을 통해 학습용 데이터를 구축하고 있다. 학습용 데이터 구축은 수집된 텍스트에 태깅 작업을 진행해야 한다. 하지만 데이터의 태깅(Tagging) 작업은 많은 인력을 필요하고 있다. 여러 논문에서는 다양한 알고리즘을 사용했으나, 본 논문에서는 일반적인 회사에서 쉽게 처리하고 최소의 인력으로 국문 텍스트 데이터에 태깅 작업을 할 수 있는 방안에 대해 연구한다.

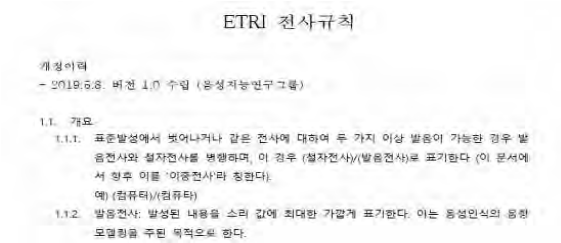
2. 본론

본 논문에서는 텍스트 기반의 훈련 데이터 구축을 위한 자동화 태깅 방법에 대해 연구를 한다. 기존에는 음성 데이터 및 영상 데이터를 속기사를 통해 전자 작업을 진행하고, 전사된 텍스트를 바탕으로 다시 태깅 작업을 진행한다. 하지만 크라우드소싱(Crowdsourcing) 방식 등으로 수집되는 대량의 데이터를 작업하기 위해서는 기존 방식으로는 한계가 있다. 본 논문에서는 한국어 기반의 자동 태깅을 할 수 있는 방안을 제시한다.

2.1 기존의 태깅 방법

전사 작업은 학습 데이터를 구축하기 위한 가장 기본적인 작업이다. 국내의 경우에는 한국전자통신연구원(ETRI)에서 제시하고 있는 ETRI 전사 규칙(그림1)을 제공받아 작업을 진행하고 있다. 속기사 업체는 별도로 규정된 작업 규칙을 사용하고 있다.

텍스트 기반의 훈련 데이터 구축을 위해서는 작업자들의 교육을 한국전자통신연구원 기반으로 진행하여 데이터를 구축한다.


ETRI 전사규칙

개정이력
- 2019.6.8. 버전 1.0 수립 (음성지능연구그룹)

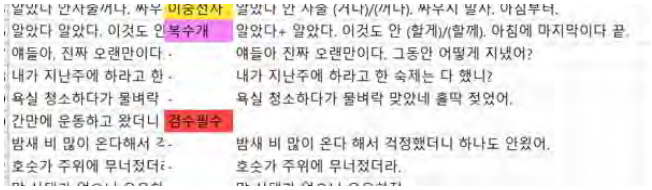
1.1. 개요
1.1.1. 표준발성에서 벗어나거나 같은 전사에 대하여 두 가지 이상 발음이 가능한 경우 발음전사와 철자전사를 병행하며, 이 경우 (철자전사)/(발음전사)로 표기한다 (이 문서에서 향후 이를 '이중전사'라 칭한다).
예) (컴퓨터)/(컴퓨타)
1.1.2. 발음전사: 발성된 내용을 소리 값에 최대한 가깝게 표기한다. 이는 음성인식의 음향 모델링을 주된 목적으로 한다.

(그림 1) 한국전자통신연구원(ETRI) 전사규칙

2.2 제안하는 태깅 방식

본 논문에서는 한글의 특성을 고려하여 100% 자동으로 태깅되는 방법이 아닌 최종 품질 검사를 통해 데이터를 검증하는 방법을 제안한다. 본 논문에서는 학습 데이터 구축을 위한 태깅 방법으로 텍스트가 된 데이터를 csv파일로 구성하여 전사규칙을 적용하는 방식을 제안한다. 본 논문에서 제시하는 것은 100% 자동화가 아닌 빠른 태깅 작업을 통해 사람이 검수를 통해 훈련 데이터 구축을 조금 더 원활히 하고자 하는 것이다.

본 논문에서 제안하는 방식은 사람이 검수를 할 수 있도록 최종 결과물을 csv 파일로 제시한다. 최종 결과물에서 사람이 수정해야 하는 부분은 검수 필수라는 단어를 제공한다.(그림 2)



(그림 2) 검수된 csv 파일에 표기된 정보

3. 실험

본 장에서는 2장에서 제안한 한국전자통신연구원(ETRI)에서 제공하는 전사규칙을 적용한다. 초기 수집한 데이터를 input 값으로 설정하고, 해당 파일을 전사규칙을 적용하여 ouput 데이터를 csv 파일로 생성하고, 적용한 전사규칙에 대해 도출하도록 수행한다. 이를 통해 텍스트 기반의 훈련 데이터 구축을 위한 자동 태깅 작업이 될 수 있는 방안을 제시한다.

3.1 실험 환경

본 논문에서는 사전에 확보한 텍스트 데이터를 csv 파일로 변환하여 한국전자통신연구원에서 제시하는 전사규칙 롤(Roll)을 적용하여 수행한다. 실험 환경은 파이썬(Python) 기반으로 개발하며 사용한 라이브러리는 (그림3)과 같다.

```
pandas==1.1.1
openpyxl==3.0.5
requests==2.24.0
```

(그림 3) 개발에 사용된 라이브러리

3.2 실험 결과 및 해석

그림 5는 본 논문에서 제시하는 방법으로 개발하여 적용된 최종 결과를 나타내는 그림이다. 그림2와 그림 4와의 차이는 단순 태깅 작업이 아닌 원문 데이터에 대한 오류를 검증하고(그림 4), 이를 위한 맞춤법 검사, 신조어, 개인정보 등에 대한 규칙을 추가하여 태깅 작업을 진행하였다. 또한 최종 파일을 csv로 할 경우 학습 데이터 구축 및 관리에 문제가 있어 데이터베이스를 설계하여 저장하는 구조로 변경하였다.

(그림 4) 오류에 대한 태깅 작업 결과

4. 결론

본 논문에서는 대용량의 학습 데이터 구축을 위한 태깅 작업을 조금 더 효율적으로 진행하기 위한 방안을 제시하였다. 한국어는 표준어와 방언, 감투어, 외래어 등이 있어 자동으로 태깅 작업을 하는 것은 많은 데이터를 통해 어렵다는 것에 결론을 내렸다. 하지만, 지속적인 학습과 형태소, 개체명 인식기 등을 활용하면 조금 더 효율을 높일 수 있을 것이라 생각한다.

앞으로 한국어 태깅 작업을 위한 별도의 한국어 사전 구축, 방언 사전 구축 등의 기본 데이터 구축이 필요하다.

본 연구에서도 향후 일반인 및 장애인들의 음성 및 텍스트 데이터를 가공하기 위해서는 초기 지식 베이스 주축이 중요하다.

Acknowledgement

본 연구는 문화체육관과부 및 한국콘텐츠진흥원의 2020년도 문화기술연구개발 지원사업으로 수행되었음

참고문헌

[1] 임해창, 임희석, 이상주, 김진동., “자연어 처리를 위한 품사 태깅 시스템의 고찰”, 한국정보과학회 정보과학회지,pp36-57, 1996.

[2] 신준철, & 옥철영. 기분석 부분 어절 사전을 활용한 한국어 형태소 분석기. 한국정보과학회 논문지, 39(5), 415 - 424. 2012.

[3] 이인근, 황도삼, 권순학, 온톨로지 구축 시스템 (An Ontology Construction System), 한글 및 한국어 정보처리 학술대회, Vol.18, p. 220-227, 2006.