|  |
| --- |
| **1. 주제**  **OCR 기술을 이용한 단어 학습 시스템**  **(가)반, 4조, 20180368, 박세진** |

|  |  |
| --- | --- |
| **2. 요약**  한국어가 아닌, 영어나 기타 언어들로 구성된 문서를 읽는 것이란 어려운 일이다. 해당 문서를 읽기 위해서는 매번 사전에 검색하는 등의 과정이 필요하다. 이전에 검색했던 단어들이 기억나지 않아 다시 동일한 단어를 검색하는 경험도 많은데, 해당 과정을 조금이나마 줄이기 위한 프로젝트이다. OCR 기술을 이용하여 다양한 형식의 파일에 있는 문장이나 단어들을 뽑아내어, 이 출력물들을 이용한 단어 학습 시스템을 제안하고자 한다. | **3. 대표 그림**    그림 1. 시스템 흐름도  OCR 기술을 이용한 변환에서 멈추는 게 아니라 거기에서 더 나아간 방법으로 해당 기술을 이용하고자 함. |

|  |
| --- |
| **4. 서론**  최신 기술이나 조금 더 자세한 정보 등을 알기 위해서는 논문 읽기가 필수라는 말이 있다. 대부분의 논문들은 우리의 모국어인 한국어가 아닌, 영어와 같은 외국어로 작성된 경우가 많다. 해당 논문들을 읽기 위해서는 단어 검색이나 문장 번역 등이 필수라고 할 수 있는데, 이 중에서도 단어 검색에 조금 더 초점을 맞춰 진행하고자 하였다.  평소 논문을 읽을 때 이전에 검색했던 단어들을 또 다시 검색하는 일이 잦았다. 이 시간이 줄어든다면 전체적인 논문 분석 시간이 단축될 수 있을 것이라고 생각하였다. 또한, 일반적으로 조금씩 연계된 내용이거나, 서로 인용을 한 논문의 경우 겹치는 단어들이 여러 번 등장하는 경우도 많다.  이러한 단어들에 대해 한 번에 완벽하게 이해하거나 암기하게 된다면 논문 뿐만 아니라 다른 기타 영어 공부에도 도움이 될 수 있을 것이라고 생각하였다. 이런 이유에 따라 다음의 단어 학습 시스템을 제안하는 바이다. |

|  |
| --- |
| **5. 본론**  1. 이미 오픈소스로 제공된 OCR 기술을 이용하여 .pdf나 .jpg 형식 등으로 이루어진 파일을 .docx나 .txt 형식의 파일로 변환하여 단어나 문장 추출에 용이하도록 만든다.  2. 변환한 파일에 있는 모든 단어들을 가져와 python의 dictionary type을 이용하여 각 단어의 빈도수 등을 확인한다.  3. 생성한 dictionary에서 관사나 영단어의 의미로 사용되고 있지 않은 단어들(약어, 사람 이름 등)을 제거하도록 한다.  4. 정제된 dictionary에서 해당 단어들의 빈도를 나타내고 있는 value 값을 확인하고 모든 단어들에 대한 순위를 정한다.  5. 특정 빈도수를 넘은 단어들을 대해 이미 구현되어 있는 machine translation model을 이용하여 자동으로 번역하도록 한다.  6. 최종적으로 번역한 단어와, 이전의 단어를 match하여 한 set로 두고, 이를 단어장 시스템에 대입하도록 한다. |

|  |
| --- |
| **6. 결론**  해당 시스템을 개발하기 위해 OCR 기술의 분석과 이 시스템에 적합한 machine translation model을 찾는 과정이 필요하다. 또한, 기술 개발에서 멈추지 않고 사용자들이 사용하는 것에 있어서 쉽고 편리할 수 있도록 하기 위한 과정 또한 필수적이다.  최종적으로 이 시스템을 이용하여 사용자들이 조금이나마 해외 문서를 더욱 더 쉽게 읽을 수 있게 될 것이다. 더 나아가 추후에는 단어 번역, 분석 과정을 이용하여 문장을 알맞게 번역하는 시스템 또한 구현할 수 있을 것이라고 기대한다. |

**7. 출처**

[1] James R. Barlow. , “OCRmyPDF”, https://github.com/jbarlow83/OCRmyPDF