지능형 영상 판독 시스템 설계를 위한 전처리 및 구현

전태현(TaeHyeon Jeon)*, 나형선(HyungSun Na)**, 안진현(Jinhyun Ahn)***, 임동혁(Dong-Hyuk Im)****
*호서대학교 컴퓨터공학과
*20161557@vision.hoseo.edu
**광운대학교 인공지능융합학과
**e-mail : nayosk@kw.ac.kr
***제주대학교 경영정보학과
***e-mail : jha@jejunu.ac.kr
****광운대학교 정보융합학부
****e-mail : dhim@kw.ac.kr

Pre-processing and implementation for intelligent imagery interpretation system

TaeHyeon Jeon *, HyungSun Na**, Jinhyun Ahn***, Dong-Hyuk Im****
* Dept. of Computer Engineering, Hoseo University
**Dept. of Applied Artificial Intelligence, Kwangwoon University
***Dept. of Management Information Systems, Jeju National University
****School of Information Convergence, Kwangwoon University

요 약

군사 분야에서 사용하는 기존 영상융합체계는 영상에서 미확인 개체를 식별하는 Activity-Based Intelligence(ABI) 기술과 객체들에 대한 지식정보를 관리하는 Structured Observation Management(SOM) 기술을 연동하여 다양한 관점에서 분석하고 있다. 그러나 군사적인 목적을 달성하기 위해서는 미래 정보가 중요하기 때문에 주변 맥락 정보를 통합하여 분석해야 할 필요성이 있으며 이를 위해 주변 맥락 정보를 분석하는 딥러닝 모델 적용이 필요하다. 본 논문에서는 딥러닝 모델 기반 영상 판독 시스템 구축을 하기 위한 전처리 과정을 설계하였다. pyhwp 라이브러리를 이용하여 영상 정보 판독 데이터를 파싱 및 전처리를 진행하여 데이터 구축을 진행하였다.

1. 서론

군사 분야에서 사용하는 기존 영상융합체계는 영상에서 미확인 개체를 식별하는 Activity-Based Intelligence(ABI) 기술과 객체들에 대한 지식정보를 관리하는 Structured Observation Management(SOM) 기술을 연동하여 영상에 대해 다양한 관점에서 분석을 하였다. 이러한 기존 기술은 영상에서 개체의 형태, 종류 등의 정적인 정보를 인식하는 데 초점을 맞추고 있다. 그러나 군사적인 목적을 달성하기 위해서는 개체들 간의 관계 및 개체가 앞으로 할 행위에 대한 정보가 더 중요하다. 따라서 개체들 간의 관계 및 개체의 미래를 추론하기 위해 주변 맥락 정보를 통합하여 분석해야 할 필요성이 있다. 주변 맥락 정보를 분석하기 위해 최근의 딥러닝 기술이 활용되어야 한다. 본 논문에서는 딥러닝 모델 구축을 위해 영상 정보 판독 시스템의 데이터를 분석하여 전처리하여 딥러닝 모델에서 사용할 수 있는 데이터 형태로 구축하는 방법을 기술한다.

2. 데이터 전처리

2.1 영상 정보 판독 업무 및 문제점

기존 영상정보 판독 업무 및 문제점의 경우 영상정보 판독관의 작성한 내용이 구조화된 형태로 관리되고 있지 않고 있다. 본 연구결과를 영상융합체계에 적용하기 위해서는 보고서의 내용을 구조화된 형태로 전처리할 필요가 있다.

2.2 hwp to txt

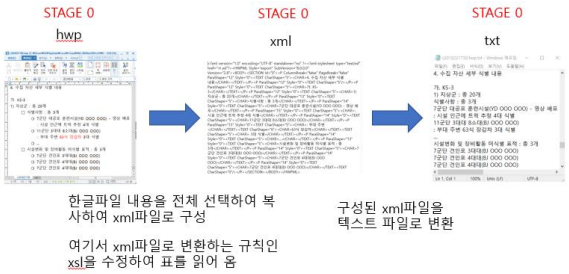
영상 판독 설계 보고서의 경우 한글파일(.hwp)로 작성되어 있다. 보고서의 내용을 쉽게 구조화하기 위해서는 한글 파일을 txt 파일 등으로의 변환이 필요하다. 본 연구에서는 이러한 문제를 해결하기 위해 pyhwp 라이브러리를 사용하였다.

2.2.1 pyhwp

pyhwp 란 HWP 문서 버전 5 파일의 내부 스트림을

분석/분리할 수 있고, 이를 Open Document format(.odt)나 일반 텍스트 문서로 변환할 수 있는 오픈소스 라이브러리이다. 그러나 pyhwp 라이브러리는 텍스트 데이터만을 파싱 하기 때문에 글 상자, 표 형식 등의 데이터는 가져오지 못하는 문제가 존재하였다. 따라서 해당 문제를 해결하기 위해 라이브러리의 변환 규칙에 대한 명세(xsl)를 수정하여 해결하였다.

그림 1 은 hwp 파일을 txt 파일로 변환하는 과정을 도식화한 것이다.



(그림 1) pyhwp 를 활용한 hwp-to-txt 과정

2.2.2 hwp to txt 데이터 클리닝

한글 파일을 txt 파일로 변환하는 과정 중 파일 깨짐 등의 현상으로 열리지 않는 파일, 암호설정으로 인한 접근 제한, 메인 데이터가 없는 경우 등의 결측치 및 이상치가 존재하는 파일은 파싱을 진행하지 않고, 삭제하였다.

2.3 데이터 클리닝

영상 판독 보고서를 분석한 결과 일정한 포맷이 존재하는 것으로 확인되었다. 그러나, 판독관이 영상 판독 보고서를 작성하는 과정에서 판독관마다 문서 작성 방식이 다르고 오타와 같은 예외도 존재한다. 이러한 데이터를 처리하기 전 선행과정으로 데이터의 특수 문자들을 제거하는 데이터 클리닝 작업을 진행하였다.

2.3.1 불일치 해결

임무현황[그림 2]의 수집자산과 수집자산 세부 식별내용[그림 3]의 수집자산이 서로 매칭된다는 것을 확인하였고, 이를 기반으로 수집 자산 세부 식별 내용의 데이터에 임무현황의 촬영지역, 촬영시간을 추가하였다.

몇몇 보고서에서 임무현황에 수집자산이 존재하지만 세부 식별내용에서는 존재하지 않는 경우, 중복되는 입수자산이 있는 경우, 순서가 바뀐 경우의 불일치가 존재하였다. 앞의 두 가지 경우는 임무현황의 데이터를 제거하였고, 마지막의 경우는 입수자산의 키워드를 기준으로 매칭시켰다.

2.3.2 결측치 대체

세부 식별 내용에서 순서는 지상군, 해군, 공군, 전략순으로 진행이 되는데, 식별사항이 없는 데이터는 삭제하였다.

2.4 데이터 분할

세부 식별 내용중 식별사항, 시설변화 및 장비활동 미 식별 표적, 구름 차폐로 식별 불가 표적 등으로 이루어져 있는데 이를 식별사항, 미 식별 표적, 그 외 식별 불가 표적으로 나누어 저장하였다.

아래의 그림 2, 그림 3 의 경우 본 연구에서 분석하는 영상 판독 보고서의 포맷 형태이다.

1. 임무현황

지도		
입수자산	KS-3(0개 궤도)	KS-3A(0개 궤도)
촬영면적/표적	0,000㎢(3.56%)/표적 00개	0,000㎢(3.56%)/표적 00개
입수영상	00씬(양호)	00씬(양호)
촬영지역	서울, 순천	인천, 의정부, 파주
촬영시간	17:00~17:10	17:10~17:20
입수시간	00:00~00:00	00:01~01:00
판독시간	00:00~00:00	01:00~02:00

(그림 2) 임무현황

4. 수집 자산 세부 식별 내용

가. KS-3

1) 지상군 : 총 20개

□ 식별사항 : 총 3개

○ 7군단 대공포 훈련시설(YD OOO OOO) - 영상 배포
: 시설 인근에 트럭 추정 4대 식별

○ 11군단 3대대 8소대(BJ OOO OOO)
: 부대 주변 63식 장갑차 3대 식별

○ …

□ 시설변화 및 장비활동 미식별 표적 : 총 3개

○ 7군단 견인포 3대대(BJ OOO OOO)

○ 7군단 견인포 4대대(BJ OOO OOO)

○ 7군단 견인포 4대대(BJ OOO OOO)

(그림 3) 수집 자산 세부 식별 내용

3. 파싱 결과

파일 이름에서 날짜, 임무현황에서 수집자산별로 촬영 시간, 촬영지역, 수집자산 세부 식별내용에서 자세한 좌표와 핵심이 되는 데이터를 매칭시켜 결과물을 tsv 파일로 변환하였다. tsv 를 사용한 이유는 촬영지역, 시간, 데이터 등에서 데이터가 여러 개일 경우 csv 의 구분 단위인 쉼표(,)를 사용하기 때문이다.

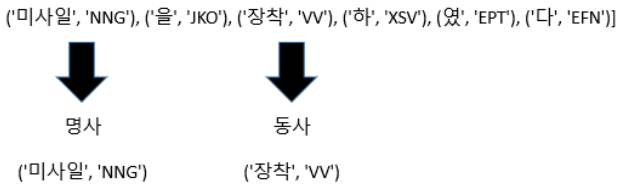
아래의 그림 4 는 본 연구에서 사용한 가공데이터를 파싱 한 결과이다.

수집자산	날짜	시간	지역	army_type	좌표data	
KS-3	20201022	17:00~17:10	서울 순천	지상군	(YD 000 000) 트럭 추정 4대 식별, (BJ 000 000) 63식 장갑차 3대 식별'	
KS-3	20201022	17:00~17:10	서울 순천	해군	(YD 000 000) PT-85 추정 12대 식별, (BJ 000 000) 자주포 1대 식별'	
KS-3	20201022	17:00~17:10	서울 순천	공군	(YD 000 000) M-2020 추정 5대 식별, (BJ 000 000) 자주포 6대 식별'	
KS-3	20201022	17:00~17:10	서울 순천	전략	(YD 000 000) 트럭 추정 0대 식별, (BJ 000 000) 자주포 6대 식별	
KS-3A	20201023	18:00~18:30	인천 의정부 파주	지상군	(YD 000 000) 트럭 추정 4대 식별, (BJ 000 000) SWAT 장갑차 3대 식별.'	
KS-3A	20201023	18:00~18:30	인천 의정부 파주	해군	(YD 000 000) BMD-1 추정 5대 식별', (BJ 000 000) 다연장로켓 1대 식별	
KS-3A	20201023	18:00~18:30	인천 의정부 파주	공군	(YD 000 000) 시설 인근에 T-70 추정 5대 식별', (BJ 000 000) M1117 6대 식별	
KS-3A	20201023	18:00~18:30	인천 의정부 파주	전략	(YD 000 000) 트럭 추정 0대 식별', (BJ 000 000) 자주포 6대 식별'	

(그림 4) tsv 결과 파일

4. 형태소 분석

본 연구에서 tsv 파일로 전처리한 데이터를 이용하여 RNN 모델 및 온톨로지를 구축하고자 한다. 이를 위해서 각 문장 데이터가 포함하고 있는 단어들이 의미적 형태를 띄어야 한다. 그러나, 한국어 텍스트는 교착어로서 구문적 복잡성을 내재하고 있고, 문법이 복잡하기 때문에 영어 자연어처리 모델을 그대로 적용할 수 없다. 따라서 본 연구에서는 형태소를 분석하여 각 어휘에서 의미적 부분이 아닌 것들을 제거한 후 명사와 동사를 분류하였다. 한국어 형태소 분석을 하기위해 꼬꼬마 형태소 분석기 라이브러리[1]를 사용하였다. 형태소 분석을 하게 될 경우 그림 5 처럼 일반명사 NNG, 동사 VV 로 분류가 된다.



(그림 5) 형태소 분석 예시

5. 향후 계획

RNN 모델 구축 및 온톨로지 구축을 위한 데이터를 수집하기 위해 hwp 문서를 pyhwp 라이브러리를 이용하여 tsv 형태로 파싱하였다. 이후, 형태소 분석기를 이용하여 단어를 얻을 수 있었다. 향후 계획으로는 온톨로지 구축을 진행하고, RNN 모델을 적용한 Sequence-To-Sequence[2] 모델과 Attention[3] 기법을 이용하여 추천 문장 생성 시스템을 구축할 예정이다.

Acknowledgement

본 연구는 국방과학연구소의 지원을 받아 수행되었음 (과제명: 지능형 행위기반 영상정보 분석기법 연구, 과제번호: UD190025FD)

참고문헌

[1] 이동주, 연종흠, 황인범, 이상구, 꼬꼬마: 관계형 데이터베이스를 활용한 세종 말뭉치 활용 도구, 2010, 정보과학지 논문지: 컴퓨팅의 실제 및 레터, Volume16, No.11, Page 1049-1050

[2] lya Sutskever, Oriol Vinyals, and Quoc V. Le. “Sequence to sequence learning with neural networks.” In Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'14). MIT Press, Cambridge, MA, USA, 3104–3112.

[3] Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio “Neural Machine Translation by Jointly Learning to Align and Translate” ICLR, 2015