Journal of the Korean Society of Marine Environment & Safety
Research Paper
Vol. 30, No. 1, pp. 097-107, February 28, 2024, ISSN 1229-3431(Print) / ISSN 2287-3341(Online) https://doi.org/10.7837/kosomes.2024.30.1.097

OCR 프로그램을 활용한 선박 항해일지 데이터 추출 모델 개발

이다인* · 김성철** · 윤익현***†

* 목포해양대학교 해상운송시스템학부 석사과정, ** 목포해양대학교 승선실습과정부 교수,
*** 목포해양대학교 항해정보시스템학부 교수

Development of a Ship's Logbook Data Extraction Model Using OCR Program

Dain Lee* · Sung-Cheol Kim** · Ik-Hyun Youn***†

* Graduate Student, Department of Maritime Transportation System, Mokpo National Maritime University, Mokpo 58628, Korea
** Professor, Division of Cadet Training, Mokpo National Maritime University, Mokpo 58628, Korea
*** Professor, Division of Navigation & Information Systems, Mokpo National Maritime University, Mokpo 58628, Korea

요 약 : 빠르게 발전하는 이미지 인식 기술에도 불구하고 표 형식의 문서와 수기로 작성된 문서를 완벽하게 디지털화하기에는 아직 어려움이 따른다. 본 연구는 표 형식의 수기 문서인 선박 항해일지를 작성하는 데에 사용되는 규칙을 이용하여 보정 작업을 수행함으로써 OCR 결과물의 정확도를 향상시키고자 한다. 이를 통해 OCR 프로그램을 통하여 추출된 항해일지 데이터의 정확성과 신뢰성을 높일 것으로 기대된다. 본 연구는 목포해양대학교 실습선 새누리호의 2023년에 항해한 57일간의 항해일지 데이터를 대상으로 OCR 프로그램 인식 후 발생한 오류를 보정하여 그 정확도를 개선하고자 하였다. 이 모델은 항해일지 기재 시 고려되는 몇 가지 규칙을 활용하여 오류를 식별한 후, 식별된 오류를 보정하는 방식으로 구성하였다. 모델을 활용하여 오류를 보정 후, 그 효과를 평가하고자 보정 전과 후의 데이터를 항차별로 구분한 후, 같은 항차의 같은 변수끼리 비교하였다. 본 모델을 활용하여 실제 셀 오류율은 약 11.8% 중 약 10.6%의 오류를 식별하였고, 123개의 오류 중 56개를 개선하였다. 본 연구는 항해일지 중 항해정보를 기입하는 Dist.Run부터 Stand Course까지의 정보만을 대상으로 수행하였다는 한계점이 있으므로, 추후 항해정보 뿐만 아니라 기상정보 등 항해일지의 더 많은 정보를 보정하기 위한 연구를 진행할 예정이다.

핵심용어 : 항해일지, 광학문자인식, 오류 인식, 오류 보정, 데이터추출모델

Abstract : *Despite the rapid advancement in image recognition technology, achieving perfect digitization of tabular documents and handwritten documents still challenges. The purpose of this study is to improve the accuracy of digitizing the logbook by correcting errors by utilizing associated rules considered during logbook entries. Through this, it is expected to enhance the accuracy and reliability of data extracted from logbook through OCR programs. This model is to improve the accuracy of digitizing the logbook of the training ship "Saenuri" at the Mokpo Maritime University by correcting errors identified after Optical Character Recognition (OCR) program recognition. The model identified and corrected errors by utilizing associated rules considered during logbook entries. To evaluate the effect of model, the data before and after correction were divided by features, and comparisons were made between the same sailing number and the same feature. Using this model, approximately 10.6% of errors out of the total estimated error rate of about 11.8% were identified, and 56 out of 123 errors were corrected. A limitation of this study is that it only focuses on information from Dist.Run to Stand Course sections of the logbook, which contain navigational information. Future research will aim to correct more information from the logbook, including weather information, to overcome this limitation.*

Key Words : *Log book, Optical Character Recognition, Error recognition, Error correction, Data extraction model*

* First Author : ekdlsdain12@gmail.com, 061-240-7281
† Corresponding Author : iyoun@mmu.ac.kr, 061-240-7283

1. 서 론

항해일지는 선박 운영에 있어서 의무적으로 비치되어야 하는 중요한 서류 중 하나이다(Park, 2022). 이 문서는 매일 선내에서 발생한 주요 사항을 상세히 기록하는 일지로, 항해당직마다 항해 및 기상 상황, 화물의 하역 작업 등을 철저히 기재한다(Jeon and Jeong, 2016). 이를 통해 선박에서의 항해, 여객 및 화물 운송 업무에 적절한 관리와 감독이 이루어졌음을 명확히 확인할 수 있다. 항해일지는 그 법적 증거능력이 높아, 재판에서 중요한 증거로 활용될 수 있다(Park, 2022; Garcia-Herrera et al., 2018; Kwon, 2014). 또한 퇴선 시에도 지참을 권고하는 만큼 이는 신뢰할 만한 기록으로 간주된다.

항해일지는 선박 운영에서 매우 중요한 역할을 수행한다. 매 당직시간이나 발생하는 사건에 따라 작성되며, 선박의 위치 또한 상세히 기록된다(Wiegmans et al., 2020). 이러한 항해일지는 그 자체로 역사적 데이터로서의 가치가 매우 높다(Seida et al., 2020). 특히 특정 일자와 해역의 상황을 유추하는 데 중요한 자료로 활용될 수 있다. 항해일지의 기록은 다양한 분야에서 연구의 대상이 되고 있다. 과거의 기후 변화를 연구하는 데 있어서 항해일지에 기록된 기상 정보와 바다의 상태 등이 매우 유용한 자료로 활용된다(Ayre et al., 2015; Wheeler, 2014; Woodruff et al., 2005; Catchpole and Faurer, 1985). 또한 17~18세기에 전 세계를 탐험한 선박들의 항적에 대한 연구에서도 항해일지는 중요한 자료로 활용되고 있다(Hong and Kim, 2020). 이와 같이 항해일지의 기록은 과거의 선박 운항 상황을 통해 다양한 분야에서의 연구 및 분석에 활용될 수 있다. 더욱 다양한 분야에서 많은 정보를 이용하기 위해서는 항해일지를 디지털화하여야 한다.

인쇄물을 디지털화하여 내용을 분석하는 시도가 활발히 이루어지고 있다. OCR(Optical Character Recognition) 기술의 발전으로 인해 정확도가 점차 향상되고 있으나, 여전히 약 80% 정도의 인식률을 갖고 있다(Kim, 2017). 특히 표 형식의 수기로 작성된 항해일지와 같은 문서의 경우 정확도가 더욱 낮다(Moon and Kim, 2023). 현재 대량의 표 형식 문서에서 정보를 추출하기 위해서는 주로 시민의 참여를 기반으로 한 전사 작업이 이루어지고 있다(Garcia-Herrera et al., 2018; Prieto et al., 2023; Lorrey et al., 2022; Teleti et al., 2023). 그러나 보다 효율적이고 신속한 디지털화를 위해서는 전사 작업이 아닌 데이터 추출 모델의 개발이 필요하다. 이를 통해 항해일지에서도 높은 정확도로 정보를 추출할 수 있게 되어 대량의 데이터를 효율적으로 활용할 수 있을 것으로 기대된다.

따라서 본 연구의 목적은 항해일지의 내용을 더욱 정확하게 디지털화하기 위하여 OCR 프로그램을 활용한 결과물을 보정하는 선박 항해일지 데이터 추출 모델을 개발하는 것이다. 이 모델은 항해일지를 작성하는 데에 사용되는 규칙을 이용하여 보정 작업을 수행함으로써 OCR 결과물의 정확도를 향상시키고자 한다. 해당 연구에서는 선박 항해일지에 기록되는 내용과 관련된 규칙을 분석하여 이를 모델링하고, OCR 프로그램을 통해 디지털화된 항해일지 데이터를 보정하는 작업을 수행하고자 한다. 이를 통해 추출된 데이터의 정확성과 신뢰성을 높일 것으로 기대된다.

2. 연구의 방법

모델 개발은 Fig. 1에 나타난 것과 같이 데이터 수집 단계부터 시작하며, 실습선 새누리호의 2023년 1년간의 항해일지 데이터 중 항해한 기간의 항해일지만 사용하였다. 데이터 수집에 활용할 OCR 프로그램을 선정하기 위하여 CER (Character Error Rate, 문자 오류율)을 기준으로 OCR 프로그램을 비교 후 선정된 프로그램을 이용하여 항해일지를 디지털화하였다.

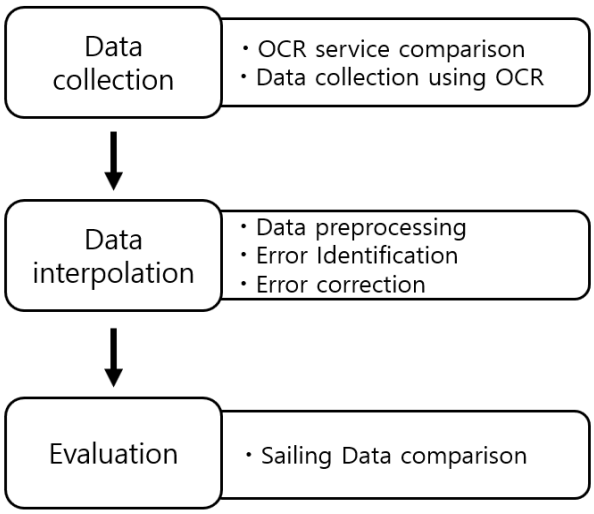

Fig. 1. Workflow of the proposed methods.

그 후, 수집된 데이터에서 오류를 식별하고 보정하기 위하여 데이터 전처리를 수행한다. 전처리가 완료된 데이터에서 본 연구에서 제안하는 규칙을 사용하여 오류를 식별한 후, 식별된 오류를 보정하였다. 실제 수집된 데이터의 오류율과 본 모델에서 식별된 오류율을 비교하고, 오류를 수정한 후 개선된 데이터의 오류율을 비교하였다.

마지막으로 오류 보정 효과를 평가한다. 오류를 보정하지 않은 OCR 프로그램 인식 후의 데이터와 본 모델을 통해 오

류를 보정한 데이터의 같은 항차, 같은 변수를 시각화하여 모델의 효과를 평가하였다.

2.1 데이터 수집

2.1.1 선박 항해일지 데이터 수집

본 연구에서는 항해일지 본란 중 기입되는 항해정보를 관련 규칙을 이용하여 오류를 식별하고 보정하고자 한다. 연구에 활용한 항해일지는 목포해양대학교 실습선 새누리호의 2023년 01월 01일부터 12월 15일까지의 항해일지로 정박한 날을 제외하여 항해한 날짜만 추출하여 총 57일의 항해일지를 활용하였다.

본 연구에서는 항해일지 본란 중 항해정보를 기입하는 Dist.Run부터 Stand Course까지의 범위에 대하여 오류를 식별하고 보정하고자 하였다. 연구 목적에 맞는 데이터를 수집하기 위하여 항해일지 좌측에 위치하는 본란(Primcipal column) 중 정오위치, 항해통계, 측정 등의 내용과 항해일지 우측에 위치한 기사란을 제외하였다.

아래의 Table 1은 본 연구에서 활용한 새누리호 항해일지의 월별 항해 일자와 월별 항해 일수, 2023년 1년간 항해한 총 일수를 나타내었다.

Table 1. Collected data list

Month.	Dates	Number of Voyage dates
Jan.	18~19	2
Feb.	22~24	3
Mar.	9~10, 20~21, 23~24	6
Apr.	12~14, 24~25, 27~28	7
May	8~12	5
June	7~9, 12~13	5
Jul.	4~5, 19~20	4
Aug.	17~20	4
Sep.	4~6, 11~14	7
Oct.	10, 12, 23~25	5
Nov.	8~9, 14~15, 26~27, 30	7
Dec.	13~14	2
Sum		**57**

2.1.2 OCR Program 비교

정확도 높은 항해일지 디지털화를 위하여 국내·외 OCR 프로그램을 비교하여 데이터 수집에 활용할 한 개의 프로그램을 선정하였다. 비교군은 웹 기반의 OCR 프로그램 중, 국내 무료 프로그램인 Naver Clova, Upstage, ePayrus TextSense와, 국외 무료 프로그램인 Google Cloud Vision, Google Keep, 국외 유료 프로그램인 AWS Textract, Microsoft Azure로 총 7개의 OCR 프로그램을 선정하였다.

인식률을 비교하기 위하여 와 같이 목포해양대학교 실습선 새누리호의 2023년 2월 24일 항해일지 본란을 활용하였다. OCR 프로그램으로 인식된 텍스트 중 프린트 된 문자를 제외하고 수기로 작성된 문자들의 CER을 기준으로 각 OCR 프로그램의 인식률을 비교하였다. Table 2는 각 OCR 프로그램의 홈페이지 주소와 각각의 프로그램을 사용하여 인식된 항해일지의 CER을 나타내었다. 인식률을 비교한 결과, AWS Textract 프로그램 활용 결과물이 CER 12.86%로 오류율이 가장 적었다. 또한 이미지에 표 형식이 있을 경우 인식 결과를 엑셀로 제공하며, 대량의 문서의 결과를 확인하기 용이한 AWS Textract를 선정하여 데이터 수집에 이용하였다. 선정된 OCR 프로그램인 AWS Textract는 아마존닷컴에서 제공하는 클라우드 컴퓨팅 플랫폼인 AWS에서 제공하는 OCR 프로그램이다.

Table 2. OCR Program list

OCR Program	Site	CER (%)
AWS Textract	https://us-east-2.console.aws.amazon.com /textract/home	12.86
ePapyrus TextSense	https://demo.epapyrus.com/ko/textsense	57.14
Google Cloud	https://cloud.google.com/vision?hl=ko	20.34
Google Keep	https://keep.google.com/	22.86
Microsoft Azure	https://documentintelligence.ai.azure.com/ studio	18.52
Naver Clova	https://clova.ai/ocr/	32.86
Upstage	https://www.content.upstage.ai/document-ai/overview	21.43

위의 과정에서 선정된 AWS Textract를 이용하여 항해일지의 이미지를 디지털화하였다. 이때, 프로그램에 업로드할 이미지는 복합기 스캔 방식을 이용하여 왜곡에 따른 인식 저하를 방지하고자 하였다(Kim et al., 2011).

Fig. 2는 스캔한 이미지의 일부로, 새누리호의 2023년 4월 13일 항해일지를 스캔한 이미지이다.

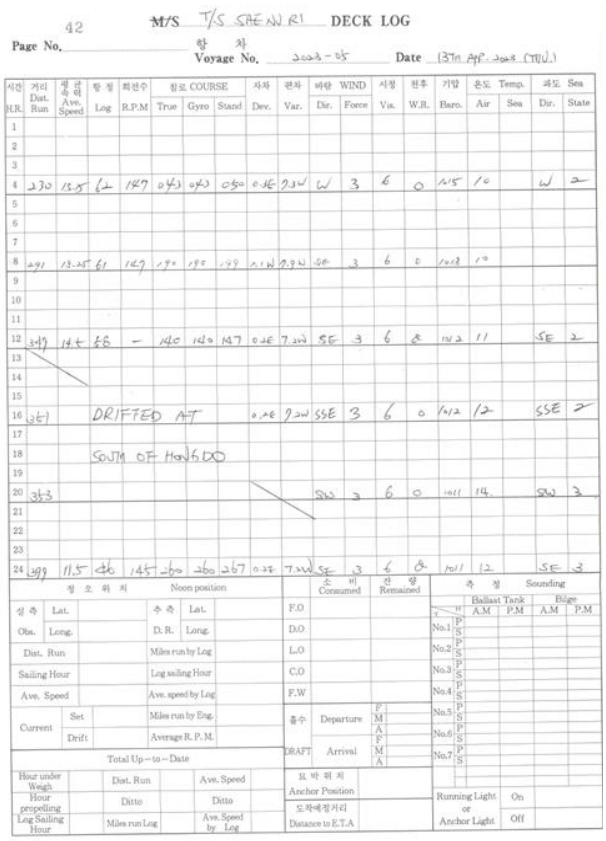

Fig. 2. Sample scan of T/S SAENURI Deck Log Book.

Fig. 3은 AWS Textract를 이용하여 항해일지를 디지털화한 엑셀 파일의 일부이다. 한글로 적혀있는‘시간’및‘평균속력’은 ‘4/1’와‘AZI’로 인식되었으며, 12시의 Ave.Speed의 경우 ‘14.5’가 ‘14.t’로 인식되었지만 항해일지의 각 시간과 정보의 칸에 맞게 인식되어진 것이 확인할 수 있다.

'4/1	'AZI	'	'	'	'	'COURSE	'
'H.R.	'Dist. Run	'Ave. Spee	'Log	'R.P.M	'True	'Gyro	'Stand
'1	'	'	'	'	'	'	'
'2	'	'	'	'	'	'	'
'3	'	'	'	'	'	'	'
'4	'230	'15.5	'62	'147	'043	'040	'050
'5	'	'	'	'	'	'	'
'6	'	'	'	'	'	'	'
'7	'	'	'	'	'	'	'
'8	'291	'15.25	'61	'147	'190	'190	'199
'9	'	'	'	'	'	'	'
'10	'	'	'	'	'	'	'
'11	'	'	'	'	'	'	'
'12	'349	'14.t	'58	'-	'140	'140	'147

Fig. 3. Result of AWS Textract recognition.

2.2 데이터 보정

2.2.1 전처리

AWS Textract를 이용하여 추출된 엑셀 파일 형태의 결과물에서 Confidence Scores를 제외하고 행은 01시부터 24시까지, 열은 Dist.Run(거리)부터 Sea.State(해상상태)까지만 추출하였다.

그 후 항해일지 본란에 기재된 시간 앞에 날짜를 추가한 후 날짜와 시간을 결합하여 연,월,일,시간이 표현된 타임스탬프를 추가하였다. 타임스탬프는 Ave.Speed의 오류를 식별하고 확인하는 과정에서 이용하며, 각 행의 위치를 잘 식별하기 위하여 추가하였다.

추출된 모든 날짜의 항해일지는 하나의 타임테이블로 통합하였으며, 항해한 날을 대상으로 데이터를 수집하였지만, 같은 날에 항해하지 않은 시간을 제외하기 위하여 Dist.Run의 기록 유무를 기준으로 하였다. 위의 기준을 적용하여 통합된 타임테이블에서 Dist.Run이 기록되지 않은 행은 제외하였다.

2.2.2 오류 식별

전처리 된 항해일지 데이터 중 OCR 프로그램을 이용하는 과정에서 오류가 발생한 셀을 식별하고자 하였다.

오류 식별 범위는 항해정보를 기입하는 Dist.Run부터 Stand Course까지였으며, 각 정보 간의 규칙을 이용하여 오류를 식별하고자 하였다.

다음과 같은 경우에 해당 셀에 오류가 있음을 식별한 후 해당 셀을 붉은색으로 표기하였다.

아래의 수식에서 오류의 유무를 확인하고자 하는 행을 (i)로, 그 이전 행을 $(i-1)$, 그 이후의 행을 $(i+1)$로 표기하였다.

① 모든 셀에 대하여 숫자와 점(.)을 제외한 문자가 있는 셀의 경우 오류로 정의하였다.

② Ave.Speed에서 새누리호의 Ship's Particular의 Service Speed 17.8(knots)에 외력을 고려하여 19(knots) 초과인 경우 오류로 정의하였다.

$$Ave.Speed(i) > 19 \quad (1)$$

③ Log에서 이전 행과 4시간 차이나는 경우, ②에서 정한 Ave.Speed의 제한범위인 19의 4배인 76 초과인 경우 오류로 정의하였다.

$$Log(i) \neq 19 \times 4 \quad (2)$$

④ True, Gyro, Stand로 표현되는 세 가지 Course에 대하여 360을 초과하는 경우 오류로 정의하였다.

$$Course(i) > 360 \quad (3)$$

이후의 규칙에서 항해를 시작하여 Dist.Run과 Log가 같은 경우에는 모두 규칙을 적용하지 않는다.

⑤ Dist.Run에서 이전 행보다 값이 작은 경우 오류로 정의하였다.

$$Dist.Run(i) < Dist.Run(i-1) \quad (4)$$

⑥ Dist.Run에서 이전 Dist.Run과 현재 Log의 합과 같지 않고, 다음 Dist.Run과 다음 Log의 차와의 차이와 같지 않을 경우, 다만 항해일지 작성 당시의 인적오류를 제외하고자 차이가 3 이하일 경우에는 오류로 정의하지 않았다.

이때, 이전 Dist.Run 또는 현재 Log가 오류로 식별되어 규칙이 적용되지 않는 경우에는 식(6)만을 적용하여 오류를 확인한다.

$$Dist.Run(i) \neq Dist.Run(i) + Log(i-1) \quad (5)$$

$$Dist.Run(i) \neq Dist.Run(i+1) - Log(i+1) \quad (6)$$

⑦ 이전 Dist.Run과 현재 Dist.Run 모두 오류가 없으며, 현재 Log가 현재 Dist.Run과 이전 Dist.Run의 차와 같지 않은 경우에 오류로 식별한다. 앞선 규칙과 마찬가지로 인적오류를 제하기 위하여 차이가 3 이하일 경우 오류로 정의하지 않았다.

$$Log(i) \neq Dist.Run(i) - Dist.Run(i-1) \quad (7)$$

⑧ 현재 Log에 오류가 없으며, 이전 행과 시간 차가 4시간인 경우 Ave.Speed가 Log를 4로 나눈 수와 같지 않을 경우 오류로 식별하였다. 인적오류를 제하기 위하여 그 차이가 0.5 이하일 경우 오류로 정의하지 않았다.

$$Ave.Speed(i) \neq Log(i) \div 4 \quad (8)$$

본 연구에서 오류로 식별하고 보정하는 범위에 포함되지 않았지만, 3가지 Course의 규칙을 정하기 위하여 Dev. 및 Var. 또한 형태가 맞지 않는 경우 오류로 식별하였다. 한 자리 숫자와 소수점 첫째 자리까지 표현되었으며, 숫자 뒤에 E 또는 W 문자가 있는 셀이 아닌 경우 오류로 정의하였다.

True Course와 Gyro Course는 자이로컴퍼스 에러만큼의 차이가 발생한다. 자이로컴퍼스는 통상적으로 그 에러를 무시할 만큼 작은 에러를 가지고 있다.

True Course와 Stand Course는 Dev.와 Var.로 인하여 차이가 발생한다. Dev.는 통상 5도 이내를 유지하도록 한다. Var.는 매년 변화가 발생하지만, 새누리호의 운항영역이 동아시아 및 한국 연안이므로 해당 지역의 Variation 분포가 10도를 넘지 않기 때문에 10도 이하를 정상범위로 설정하였다.

Stand Course에서 True Course로 침로의 개정을 시행할 때, Dev.와 Var.가 WEST일 경우 그 값만큼 Stand Course에서 빼주고, EAST일 경우 그 값만큼 Stand Course에서 더해준다. 본 연구에서는 WEST일 경우 양수로, EAST일 경우 음수로 치환한 후 Stand Course에서 True Course를 계산할 때 Dev.와 Var.의 합을 빼주었다.

이후의 규칙은 3개의 Course 모두 오류가 없다는 것을 전제로 적용한다. 만약, 같은 행에 한 개 이상의 Course가 오류인 경우 규칙을 적용하지 않는다.

⑨ 인적오류를 포함하여 True Course와 Gyro Course의 차이가 3도 이상일 경우, Stand Course와의 차이가 Dev.와 Var.로 인한 차이인 15도 이상인 방위를 오류로 정의한다.

$$|True\ Course(i) - Gyro\ Course(i)| > 3 \quad (9)$$

$$|Error\ Course(i) - Stand\ Course(i)| > 15 \quad (10)$$

(Error Course는 True Course 또는 Gyro Course)

⑩ True Course와 Gyro Course가 일치하는 경우 True Course와 Stand Course가 15도 이상인 경우 Stand Course가 오류임을 정의한다.

$$True\ Course(i) = Gyro\ Course(i) \quad (11)$$

$$|True\ Course(i) - Stand\ Course(i)| > 15 \quad (12)$$

⑪ 같은 행의 Dev.와 Var. 모두 오류가 없으며 True Course와 Gyro Course가 일치하는 경우, True Course와 Stand Course의 차이가 Dev. 및 Var.의 합과 1도 이상 차이 나는 경우 Stand Course가 오류임을 정의한다.

$$Diff = |True\ Course(i) - Stand\ Course(i)| \quad (13)$$

$$DV = |Dev.(i) + Var.(i)| \quad (14)$$

$$|Diff - DV| > 1 \quad (15)$$

(Diff = True Course와 Gyro Course의 차,
DV = Dev.와 Var.의 합)

2.2.3 오류 보정

2.2.2에서 식별한 오류를 식별한 규칙과 비슷한 규칙을 적용하여 보정하였다.

오류로 식별된 Dist.Run을 보정한 규칙은 다음과 같다.

① 이전 행과의 시간 차가 4시간이며 이전 Dist.Run과 현재 Log 모두 오류가 없는 경우 식(16)를 통하여 보정한다.

$$Dist.Run(i) = Dist.Run(i-1) + Log(i) \quad (16)$$

② 이전 행과의 시간 차가 4시간이며 이전 Dist.Run과 현재 Ave.Speed는 오류가 없지만 현재 Log는 오류가 있는 경우 식(17)을 통하여 보정한다.

$$Dist.Run(i) = Dist.Run(i-1) + Ave.Speed(i) \times 4 \quad (17)$$

③ 이전 행과의 시간 차가 24시간 이상일 경우 항해의 시작으로 간주하여 현재 Log에 오류가 없을 경우 식(18)을 이용하여 오류를 보정한다. 만일 Log에 오류가 있고 다음 Dist.Run과 Log에 오류가 없을 경우에는 식(19)을 이용하여 보정한다.

$$Dist.Run(i) = Log(i) \quad (18)$$

$$Dist.Run(i) = Dist.Run(i+1) - Dist.Run(i+1) \quad (19)$$

오류로 식별된 Ave.Speed를 보정한 규칙은 다음과 같다.

④ 이전 행과의 시간 차가 4시간이며, 현재 Log에 오류가 없을 경우 식(20)을 이용하여 보정한다.

$$Ave.Speed(i) = Log(i) \div 4 \quad (20)$$

오류로 식별된 Log을 보정하는 규칙은 다음과 같다.

⑤ 현재 Dist.Run과 이전 Dist.Run 모두 오류가 없을 경우 식(21)을 이용하여 보정한다.

$$Log(i) = Dist.Run(i) - Dist.Run(i-1) \quad (21)$$

⑥ 이전 행과의 시간 차가 4시간이며, 현재와 이전 두 개의 Dist.Run 중 하나라도 오류가 있고 Ave.Speed에 오류가 없을 경우 식(22)을 이용하여 보정한다.

$$Log(i) = Ave.Speed(i) \times 4 \quad (22)$$

⑦ RPM에서 오류가 난 경우 이전 RPM에 오류가 없다면 이전 RPM로 보정하였다.

Course를 보정하기 위하여 오류로 식별된 Dev.와 Var.의 경우 다음과 같은 규칙을 이용하여 보정한 후 Course 보정에 활용되었다.

⑧ 선박의 선수방위에 따라 변화하는 Dev.의 경우 오류로 식별된 Dev.와 True Course가 같거나 비슷한 경우 중 오류로 식별되지 않은 Dev.의 값으로 보정하였다.

⑨ 위치에 따라 변화하는 Var.의 경우, 오류로 식별된 Var.와 가장 가까운 행 중 오류로 식별되지 않은 Var.의 값으로 보정하였다.

Course를 보정한 규칙은 다음과 같다.

⑩ True Course와 Gyro Course 중 하나만 오류로 식별되었다면, 오류로 식별되지 않은 나머지 값으로 보정하였다.

⑫ True Course와 Gyro Course 모두 오류로 식별되었으며 Stand Course는 오류로 식별되지 않은 경우 식(23)을 이용하여 오류를 보정하였다.

$$Error\ Course(i) = Stand\ Course(i) - Dev.(i) - Var.(i) \quad (23)$$

(Error Course = True Course & Gyro Course)

⑬ Stand Course가 오류로 식별된 경우, 식(24)를 이용하여 오류를 보정하였다.

$$Stand\ Course(i) = True\ Course(i) + Dev.(i) - Var.(i) \quad (24)$$

아래의 Table 3과 Table 4는 오류 식별 및 보정 시 서로 관련이 있는 정보들을 한 table에 기입하였으며, 본 모델에서 제안하는 규칙을 항해일지 형태의 table에 요약한 것이다. 좌측부터 우측으로, 상단에서 하단의 순서로 알파벳을 지정하였다.

Table 3. Rules for identified error (Dist.Run, Ave.Speed, Log)

H.R.	Dist.Run	Ave.Speed ≤19	Log ≤76
4	ⓐ = ⓓ - ⓕ	ⓑ	ⓒ
8	ⓓ = ⓐ + ⓕ	ⓔ = ⓕ / 4	ⓕ = ⓔ × 4

(ⓐ = 04시 Dist.Run, ⓑ = 04시 Ave.Speed, ⓒ = 04시 Log, ⓓ = 08시 Dist.Run, ⓔ = 08시 Ave.Speed, ⓕ = 08시 Log)

Table 4. Rules for identified error (Courses, Dev., Var.)

Course ≤360			Dev. ≤5	Var. ≤10
True	Gyro	Stand		
ⓖ = ⓗ	ⓗ = ⓖ	ⓘ = ⓖ + (ⓙ+ⓚ)	ⓙ	ⓚ
ⓖ = ⓘ - (ⓙ+ⓚ)	ⓗ = ⓘ - (ⓙ+ⓚ)	ⓘ = ⓗ + (ⓙ+ⓚ)	(Positive number if W, If E is negative number)	

(ⓖ = True Course, ⓗ = Gyro Course, ⓘ = Stand Course, ⓙ = Deviation, ⓚ = Variation)

2.3 모델 평가

본 모델을 통한 항해일지 디지털화 과정에서 발생한 오류 보정 효과를 평가하기 위하여 보정 전 데이터와 보정 후 데이터를 비교하였다.

데이터를 비교하기 위하여 오류 식별 및 보정을 위하여 통합하였던 타임테이블을 항차별로 구분하였다. 항차별로 구분하기 위하여 Dist.Run과 Log의 값이 같을 경우를 항차의 시작으로 정하였다. 보정 후 데이터의 경우, 위의 기준을 적용하였을 때 항차별로 잘 구별되었다. 그러나 보정 전 데이터의 경우, Dist.Run과 Log 값에 오류가 있어 그 값이 같은지 확인이 불가능한 경우에는 잘 적용되지 않았다. 보정 전 데이터를 잘 구분하기 위하여, 위의 기준과 함께 이전 행과 시간 차이가 24시간 즉, 하루 이상 차이가 있는 경우에는 항차의 시작으로 간주하였다. 항차별로 분류한 보정 전 데이터와 보정 후 데이터를 같은 항차끼리, 같은 변수를 비교하였다.

3. 결과 및 토의

3.1 데이터 보정

3.1.1 전처리

오류를 식별하고 보정하기 위하여 AWS Textract 인식 후 제공된 결과물 중 분석에 불필요한 Confidence Scores를 제외하고 행은 01시부터 24시까지, 열은 Dist. Run (거리)부터 Sea. State (해상상태)까지만 추출하였다.

그 후 오류 식별 및 보정 시 활용하기 위하여 시간열 앞에 날짜를 추가하여 날짜와 시간이 표현된 타임스탬프를 추가하였다. Fig. 4는 타임스탬프를 추가한 타임테이블 중 일부이다.

추출된 모든 날짜의 항해일지를 하나의 타임테이블로 통합하였으며, 해당 데이터 중 항해하지 않은 시간대를 제외하기 위하여 Dist.Run이 기록되어 있지 않은 행을 제외하였

다. Fig. 5는 하나의 타임테이블로 데이터를 통합한 후, Dist.Run이 기록되지 않은 행을 제외한 타임테이블의 일부이다.

DateTime	1 DistRun	2 AveSpeed	3 Log
2023-10-24 01:00:00	''	''	''
2023-10-24 02:00:00	''	''	''
2023-10-24 03:00:00	''	''	''
2023-10-24 04:00:00	'131'	'2.00'	'8'
2023-10-24 05:00:00	''	''	''
2023-10-24 06:00:00	''	''	''
2023-10-24 07:00:00	''	''	''
2023-10-24 08:00:00	'156'	'1.3'	'25'
2023-10-24 09:00:00	''	''	''
2023-10-24 10:00:00	''	''	''
2023-10-24 11:00:00	''	''	''
2023-10-24 12:00:00	'213'	'14.3'	'57'

Fig. 4. Example Log Book timetable.

DateTime	1 DistRun	2 AveSpeed	3 Log	4 RPM
2023-01-18 16:00:00	"45"	"16.0"	"45"	"145"
2023-01-18 20:00:00	"102"	"14.5"	"58"	"145"
2023-01-19 00:00:00	"163"	"14.75"	"60"	"145"
2023-01-19 04:00:00	"228"	"16.3"	"65"	"145"
2023-01-19 08:00:00	"290"	"15.5"	"62"	"14t"
2023-02-22 16:00:00	"31"	"15.5"	"31"	"146"
2023-02-22 20:00:00	"90"	"148"	"59"	"145"
2023-02-23 00:00:00	"148"	"14.50"	"58"	"145"
2023-02-23 04:00:00	"212"	"16"	"64"	"146"
2023-02-23 08:00:00	"276"	"16.0"	"64"	"146"
2023-02-23 12:00:00	"336"	"15.00"	"60"	"145"
2023-02-23 16:00:00	"386"	"125"	"50"	"145"
2023-02-23 20:00:00	"422"	"9.0"	"36"	"-"
2023-02-24 04:00:00	"474"	"13.0"	"52"	"146"
2023-02-24 08:00:00	"335"	"15.3"	"61"	"146"
2023-03-09 16:00:00	"44"	"15.7"	"44"	"145"
2023-03-09 20:00:00	"108"	"16.0"	"I"	"145"

Fig. 5. Combined Logbook Timetable.

3.1.2 오류 식별

항해일지 기입 시 고려되는 사항을 바탕으로 2.2.2에서 정한 규칙을 이용하여 항해일지를 AWS Textract를 활용하여 인식한 후 발생한 오류를 식별하였다. 오류가 발생한 셀의 위치를 쉽게 인식할 수 있도록 Fig. 6과 같이 오류가 발생한 셀을 붉은색으로 표시하였다.

DateTime	DistRun	AveSpeed	Log	RPM
2023-01-18 16:00:00	45	16.0	45	145
2023-01-18 20:00:00	102	14.5	58	145
2023-01-19 00:00:00	163	14.75	60	145
2023-01-19 04:00:00	228	16.3	65	145
2023-01-19 08:00:00	290	15.5	62	14t
2023-02-22 16:00:00	31	15.5	31	146
2023-02-22 20:00:00	90	148	59	145
2023-02-23 00:00:00	148	14.50	58	145
2023-02-23 04:00:00	212	16	64	146
2023-02-23 08:00:00	276	16.0	64	146
2023-02-23 12:00:00	336	15.00	60	145
2023-02-23 16:00:00	386	125	50	145

Fig. 6. Visualized Identified Errors.

DateTime	DistRun	AveSpeed	Log	RPM
2023-01-18 16:00:00	45	16.0	45	145
2023-01-18 20:00:00	102	14.5	58	145
2023-01-19 00:00:00	163	14.75	60	145
2023-01-19 04:00:00	228	16.3	65	145
2023-01-19 08:00:00	290	15.5	62	14t
2023-02-22 16:00:00	31	15.5	31	146
2023-02-22 20:00:00	90	148	59	145
2023-02-23 00:00:00	148	14.50	58	145
2023-02-23 04:00:00	212	16	64	146
2023-02-23 08:00:00	276	16.0	64	146
2023-02-23 12:00:00	336	15.00	60	145
2023-02-23 16:00:00	386	125	50	145

Fig. 7. Visualization of Actual Errors.

식별된 110개의 오류 중 52개가 숫자이지만 숫자 외의 문자로 인식한 경우로 식별된 오류 중 약 47%를 차지하였다. 그 외에도 Ave.Speed의 경우 소수점을 인식하지 못하거나 다른 문자로 인식한 경우, 숫자를 다른 숫자로 인식한 경우에도 오류로 식별되었다.

본 모델이 오류를 정확하게 식별하였는지 확인하기 위하여 실제 오류 식별에 활용된 항해일지 데이터와 AWS Textract 인식 후의 데이터를 비교하여 오류가 있는 부분을 확인하였다. Fig. 7은 실제 데이터와 비교 후 확인된 오류를 시각화한 테이블의 일부이다.

본 모델에서는 총 1038개의 셀 중 110개의 오류를 식별하였는데, 실제 항해일지 데이터와 비교한 결과 AWS Textract로 인식된 데이터에서는 123개의 오류가 발생하였다. 본 모델로 측정한 AWS Textract를 통하여 추출된 데이터의 셀 오류율은 약 10.6%였으며, 실제 셀 오류율은 약 11.8%로 약 1%의 차이를 보이며 대부분의 오류를 식별하였음을 확인하였다.

아래의 Table. 5는 본 모델이 식별한 오류 개수와 실제 AWS Textract 인식 결과의 오류 개수와 두 경우에서의 오류율을 나타내었다.

Table 5. Number of Errors

Comparative method	Total number of cells	Number of error cells	Persentage (%)
Proposed model	1038	110	10.6
Direct observation	1038	123	11.8

3.1.3 오류 보정

2.2.3에서 정한 규칙을 이용하여 오류를 보정한 후, 보정의 정확도를 확인하기 위하여 Fig. 8과 같이 본 모델을 통하여 보정된 데이터를 실제 항해일지 데이터와 비교하였다. 실제 항해일지와 비교한 AWS Textract 인식 후 실제 오류 개수를 기준으로, 123개의 오류 중 56개를 개선하여 67개의 오류가 확인되었다.

아래의 Table 6은 실제 AWS Textract 인식 결과 오류율과 모델을 통해 보정한 후의 오류율을 비교하였다.

Table 6. Before and after correction

	Total number of cells	Number of error cells	Persentage (%)
Before	1038	123	11.8
After	1038	67	6.5

보정하지 못한 오류에 대하여 실제 항해일지 내용과 비교한 결과, Fig. 9과 Fig. 10의 경우와 같이 새누리호 항해일지에서 Ave.Speed 기재 시 소수점 첫째 자리까지 표현하는 방식과 소수점 둘째 자리까지 표현하는 방식을 혼용하여 이에 따른 오차가 발생한 경우가 3건이었다. 본 모델에서는 보정된 Ave.Speed를 소수점 둘째 자리까지 표현하였다.

또한 원래 항해일지를 기입할 당시 당직사관의 계산 실수로 인하여 잘못 기입된 경우 규칙에 맞춰 계산한 값과 맞지 않은 경우가 3건이었다. Fig. 11과 Fig. 12의 경우 11월 30일 20시의 Log를 제외하고 나머지 값은 본 모델을 통하여 보정한 값과 실제 항해일지에 기입된 값이 모두 동일하였다. 실제 항해일지에 기입된 값을 이용하여 Log를 계산해보면, 이전 Dist.Run과의 차, Ave.Speed를 4배한 값 모두 49임을 알 수 있다.

DateTime	DistRun	AveSpeed	Log
2023-03-10 04:00:00	233	15.8000	63
2023-03-10 08:00:00	295	15.5000	62
2023-03-20 16:00:00	40	16.7000	40
2023-03-20 20:00:00	102	15.5000	62
2023-03-21 00:00:00	162	15.0000	60
2023-03-21 04:00:00	222	15.2500	61
2023-03-21 08:00:00	284	15.5000	62
2023-03-23 16:00:00	54	16.4000	54
2023-03-23 20:00:00	116	15.5000	62
2023-03-24 00:00:00	176	15.0000	60
2023-03-24 04:00:00	229	13.2500	53
2023-03-24 08:00:00	281	13.0000	52
2023-04-12 16:00:00	50	15.2000	50
2023-04-12 20:00:00	108	14.5000	58
2023-04-13 00:00:00	168	15.0000	60
2023-04-13 04:00:00	230	15.5000	62
2023-04-13 08:00:00	291	15.2500	61
2023-04-13 12:00:00	349	14.5000	58

Fig. 8. Error Visualization Post-correction.

DateTime	DistRun	AveSpeed	Log
2023-02-22 20:00:00	90	14.7500	59

Fig. 9. 20:00 22th Feb. Post-Correction Ave.Speed.

1 DateTime	2 DistRun	3 AveSpeed	4 Log
2023-02-22 20:00:00	90	14.8000	59

Fig. 10. 20:00 22th Feb. Actual Ave.Speed.

DateTime	DistRun	AveSpeed	Log
2023-11-30 16:00:00	89	15.0000	60
2023-11-30 20:00:00	138	12.3000	49

Fig. 11. 30th Nov. Post-Correction Dist.Run & Log

1 DateTime	2 DistRun	3 AveSpeed	4 Log
2023-11-30 16:00:00	89	15	60
2023-11-30 20:00:00	138	12.3000	47

Fig. 12. 30th Nov. Actual Dist.Run & Log.

3.2 모델 평가

모델의 보정 효과를 평가하기 위하여 보정 전 데이터와 보정 후 데이터를 비교하기 위하여 오류 식별 및 보정을 위하여 통합했던 타임테이블을 항차별로 구분하였다. 2.3에서

언급한 기준으로 구분한 결과, 보정 전과 후 데이터 모두 실제 항해한 날짜별로 잘 구분되었다. Fig. 13의 경우 항차별로 구분한 데이터 중 2월 22일부터 24일까지 항해한 항차의 보정 전 데이터를 나타내고, Fig. 14의 경우 같은 항차의 보정 후 데이터를 나타낸다.

DateTime	1 DistRun	2 AveSpeed	3 Log
2023-02-22 16:00:00	"31"	"15.5"	"31"
2023-02-22 20:00:00	"90"	"148"	"59"
2023-02-23 00:00:00	"148"	"14.50"	"58"
2023-02-23 04:00:00	"212"	"16"	"64"
2023-02-23 08:00:00	"276"	"16.0"	"64"
2023-02-23 12:00:00	"336"	"15.00"	"60"
2023-02-23 16:00:00	"386"	"125"	"50"
2023-02-23 20:00:00	"422"	"9.0"	"36"
2023-02-24 04:00:00	"474"	"13.0"	"52"
2023-02-24 08:00:00	"335"	"15.3"	"61"

Fig. 13. Pre-correction of sailing 22nd~24th, Feb.

DateTime	1 DistRun	2 AveSpeed	3 Log
2023-02-22 16:00:00	31	15.5000	31
2023-02-22 20:00:00	90	14.7500	59
2023-02-23 00:00:00	148	14.5000	58
2023-02-23 04:00:00	212	16	64
2023-02-23 08:00:00	276	16	64
2023-02-23 12:00:00	336	15	60
2023-02-23 16:00:00	386	12.5000	50
2023-02-23 20:00:00	422	9	36
2023-02-24 04:00:00	474	13	52
2023-02-24 08:00:00	535	15.3000	61

Fig. 14. Post-correction of sailing 22nd~24th, Feb.

Fig. 15의 경우 새누리호의 항해일지 중 11월 14일부터 15일까지의 데이터 중 Dist.Run을 계단상그래프를 통하여 비교한 것이다. 붉은색의 선으로 표현한 보정 전 데이터에서 14일 16시의 Dist.Run이 숫자가 아닌 문자로 인식되어 그래프에 표현이 되지 않았으며, 14일 20시의 Dist.Run이 본래 값인 '148'이 아닌 '14'로 인식되어 보정 후의 값과 큰 차이를 보였다. 이 두 값 모두 본래 값으로 보정되어 파란색 선으로 표현한 보정 후 데이터는 4시간의 당직시간동안 비슷한 간격을 유지하며 증가하는 것을 확인할 수 있다.

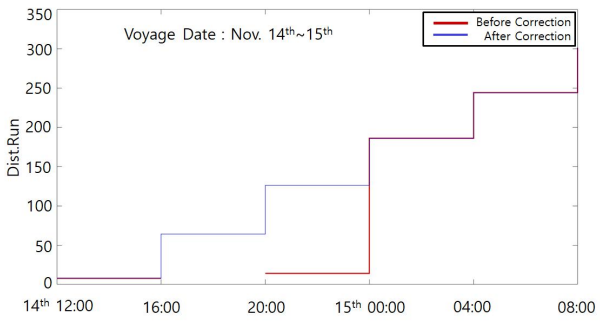

Fig. 15. Comparing Dist.Run Before and After Correction.

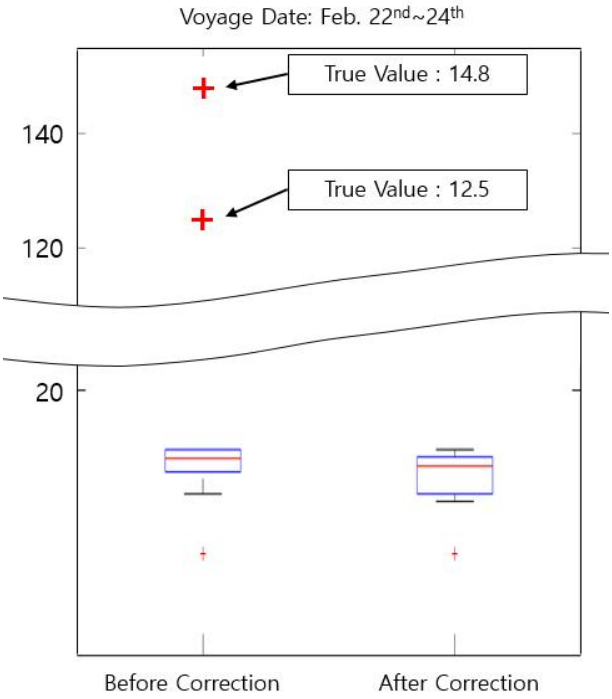

Fig. 16. Comparing Ave.Speed Before and After Correction.

Fig. 16의 경우 새누리호의 항해일지 중 2월 22일부터 24일까지의 데이터 중 Ave.Speed를 박스플롯을 통하여 비교한 것이다. 해당 항차의 항해동안 기록된 10개의 Ave.Speed 중 2개를 제외한 나머지 값은 오류가 없이 잘 인식되어, IQR (Inter Quatile Range)의 위치는 보정 전과 후 데이터에서 비슷하였다. 이를 통하여 실습선 새누리호의 해당 항차 당시 최솟값보다 작은 이상치 '9'를 제외하고는 비슷한 속력을 유지한 것을 확인할 수 있다. 그러나 보정 전 데이터의 경우 Ave.Speed의 소수점을 인식하지 못하여 '148'과 '125'로 인식된 것이 최댓값보다 큰 이상치로 확인되었다.

계단상그래프와 박스플롯을 이용한 보정 전과 후의 데이터 비교를 통하여 보정이 효과적으로 이루어졌다는 것을 확인할 수 있다.

3.3 결과에 대한 토의

항해일지는 그 역사가 적어도 18세기 초까지 거슬러 올라가야 할 만큼 역사적인 기록물이다(Garcia-Herrera et al., 2018). 영국의 경우 18세기부터 영국 해군이 작성한 항해일지를 영국 국립문서보관소에 보관하고 있다(Ayre et al., 2015). 오랜 역사를 지니고있는 항해일지는 항해 시 위치와 기상상황, 항해상황에 대하여 기록되어 있으므로 과거의 기상 정보와 해양에 관한 연구, 과거 무역의 흐름 등 다양한 연구에 활용될 수 있다(Lorrey et al., 2022; Wiegmans et al., 2020).

최근 이미지를 인식하여 텍스트로 변환시켜 주는 기술이 매우 발전하고 있지만, 표 형식의 문서의 경우 표의 형태에 따라 같은 열이나 행이라도 OCR 프로그램 인식 시 다르게 배열될 수 있다(Prieto et al., 2023). 수기로 작성된 문서 또한 작성한 사람마다 필체가 다르기 때문에 인식하기에 어려움이 있다. 항해일지는 이 두 특징이 모두 있는 문서이므로, 디지털 데이터로 변환하기에 어려움이 따른다.

선행 연구에서는 주로 crowd sourcing으로 항해일지 데이터를 디지털화하여 연구를 진행하였다(Garcia-Herrera et al., 2018; Prieto et al., 2023; Lorrey et al., 2022; Teleti and Wood, 2023). 본 연구에서는 대중의 참여 없이도 대량의 항해일지를 디지털화하고자 하였다. 이를 위하여 항해일지의 정보를 기입할 때 참고되는 몇 가지 규칙들을 이용하여 OCR 프로그램으로 추출된 항해일지 데이터에서 프로그램 인식 시 생성된 오류를 식별한 후 보정하는 모델을 개발하였다. 항해일지의 정보 중 서로 연관이 있는 정보를 이용하여 오류를 수정하므로 오류 없이 제대로 인식된 정보를 이용하여 오류를 수정하여 OCR 프로그램 인식 결과를 더욱 개선시킬 수 있었다.

오류를 보정한 후, 통합된 타임테이블을 항차별로 나누어 같은 항차의 같은 변수를 비교하여 모델의 보정 효과를 평가하였다. 11월 중 항해한 항차의 Dist.Run을 비교한 계단상 그래프와 2월 중 항해한 항차의 Ave.Speed를 비교한 박스플롯을 확인한 결과, 두 경우 모두 오류가 제대로 보정되었음을 확인할 수 있었다.

본 연구에서는 항해정보를 기입하는 란인 Dist.Run부터 Stand Course까지의 범위에서 오류를 식별하고 보정하였다. 해당 모델을 이용하여 AIS 정보가 남아있지 않은 과거 선박의 항해일지를 디지털화한다면, 해당 선박의 항해거리와 침로 등을 이용하여 항로를 추정할 수 있고, 항해거리, 회전수, 평균속력 등을 이용하여 당직시간별 엔진의 효율을 추정하는 등 당시 항해상황을 이해하는데 기여할 수 있을 것이다.

4. 결 론

빠르게 발전하는 이미지 인식 기술에도 불구하고 표 형식의 문서와 수기로 작성된 문서를 완벽하게 디지털화하기에는 아직 어려움이 따른다. 본 연구의 목적은 두 가지 특성을 모두 가진 항해일지를 항해일지 기재 시 고려되는 몇 가지 규칙을 활용하여 OCR 프로그램 인식 후 오류가 발생한 부분을 식별하고 보정하고자 하였다.

본 연구에서는 목포해양대학교 실습선 새누리호의 2023년 중 항해한 57일의 항해일지를 대상으로 AWS Textract를 이용하여 디지털화한 것을 몇 가지 규칙을 활용하여 오류를 식별하고 보정하였다. 연구에서 활용된 규칙은 항해일지 기재 시 사용하는 규칙을 기반으로 형성되었다. 본 모델의 보정 효과를 평가하기 위하여 보정 전과 후의 데이터를 항차별로 구별한 후, 같은 항차의 같은 변수를 비교한 후 시각화하였다.

본 연구의 한계점으로는, 다른 셀과의 관계를 통한 규칙을 이용하였기 때문에 오류를 식별하고자 하는 셀과 관련이 있는 모든 셀의 값이 오류로 식별된 경우, 해당 셀의 오류 유무를 확인하지 못하여 보정할 수 없었다는 점이다. 또한, 항해일지 중 항해정보를 기입하는 Dist.Run부터 Stand Course까지의 정보만을 대상으로 수행하였다는 점이다. 특정 규칙에 의하여 기입되는 것이 아닌 당시 상황을 그대로 기입하는 기상정보에 경우 그것에 특화된 오류 보정 모델이 필요할 것으로 예상된다.

추후 항해정보 뿐만 아니라 앞서 말한 기상정보 등 항해일지의 더 많은 정보를 보정하고, 기사란에 기재된 당직시간 당 기재한 선박의 위치를 활용하기 위하여 연구를 진행할 예정이다. 이를 통해 기상청 정보가 남아있지 않은, 혹은 기상청 정보가 있는 시기라도 선박과 거리가 있는 기상 관측부이가 아닌 선박의 위치에서의 기상정보를 해당 선박의 항해정보와 결합하여 선박의 항해와 기상에 관한 연구에 활용할 수 있을 것으로 예상된다.

References

[1] Ayre, M., J. Nicholls, C. Ward, and D. Wheeler(2015), Ships' logbooks from the Arctic in the pre instrumental period, Geoscience Data Journal, Vol. 2, No. 2, pp. 53-62.

[2] Catchpole, A. J. W. and M. A. Faurer(1985), Ships' log-books, sea ice and the cold summer of 1816 in Hudson Bay and its approaches, Arctic, pp. 121-128.

[3] García-Herrera, R., D. Barriopedro, D. Gallego, J. Mellado-Cano, D. Wheeler, and C. Wilkinson(2018), Understanding

weather and climate of the last 300 years from ships' logbooks, Wiley Interdisciplinary Reviews: Climate Change, Vol. 9, No. 6, pp. e544.

[4] Hong, O. S. and N. H. Kim(2020), The Exploratory Voyages of Joseon by Europeans around the 19th Century and the Records of Voyages, Journal of the Center for Korean Studies, Inha University, Vol. 58, pp. 9-39.

[5] Jeon, J. H. and T. G. Jeong(2016), Studies on the Improvement and Analysis of Data Entry Error to the AIS System for the Traffic Ships in the Korean Coastal Area, The Journal of Fisheries and Marine Sciences Education, Vol. 28, No. 6, pp. 1812-1821.

[6] Kim, D. Y., H. S. Kim, J. S. Kim, S. C. Kim, and K. I. Hwang(2011), Development of a TTS based Book Reader for the Blind, Annual Spring Conference of KIPS, Vol. 18, No. 2, pp. 422-424.

[7] Kim, J. H.(2017), A Case Study of Transcription Programs Based on Citizens' Contribution to Overseas Archival Institutions, Journal of Korean Society of Archives and Records Management, Vol. 17, No. 4, pp. 51-86.

[8] Kwon, O.(2014), A Study on the salvage of shipwreck vessel and the reward of salvage charge, The Journal of Korea Research Society for Customs, Vol. 15, No. 4, pp. 239-259.

[9] Lorrey, A. M., P. R. Pearce, R. Allan, C. Wilkinson, J. M. Woolley, E. Judd, S. Mackay, S. Rawhat, L. Slivinski, S. Wilkinson, E. Hawkins, P. Quesnel, and G. P. Compo(2022), Meteorological data rescue: Citizen science lessons learned from Southern Weather Discovery, Patterns, Vol. 3, No. 6.

[10] Moon, S. H. and J. W. Kim(2023), Deep Learning-based Automated Sentence Segmentation for Digitization of Offline Document from Industrial Jobsites, 2023 Spring Joint Conference of KORMS and KIIE, pp. 3090-3097.

[11] Park, Y. S.(2022), A Proposed Amendment to the Korean Seafarers' Act on Log Book Entries, The Journal of Korea Maritime Law Association, Vol. 44, No. 3, pp. 227-264.

[12] Prieto, J. R., J. Andrés, E. Granell, J. A. Sánchez, and E. Vidal(2023), Information extraction in handwritten historical logbooks, Pattern Recognition Letters, Vol. 172, pp. 128-136.

[13] Seida, K., H. Chiba, and A. Ohsaka(2020), Digitalizing the Full Documents of Deck Log Book of Sail Training Ship "Kaiwo Maru I" and a Study for the Usage of the Data, The Journal of Japan Institute of Navigation, Vol. 143.

[14] Teleti, P., E. Hawkins, and K. R. Wood(2023), Digitizing weather observations from World War II US naval ship logbooks, Geoscience Data Journal.

[15] Wheeler, D.(2014), Hubert Lamb's 'treasure trove': ships' logbooks in climate research, Weather, Vol. 69, No. 5, pp. 133-139.

[16] Wiegmans, B., P. Witte, M. Janic, and T. de Jong(2020), Big data of the past: Analysis of historical freight shipping corridor data in the period 1662-1855, Research in Transportation Business & Management, Vol. 34, pp. 100459.

[17] Woodruff, S. D., H. F. Diaz, S. J. Worley, R. W. Reynolds, and S. J. Lubker(2005), Early ship observational data and ICOADS, Climatic Change, Vol. 73, No. 1-2, pp. 169-194.

Received : 2024. 02. 05.
Revised : 2024. 02. 22.
Accepted : 2024. 02. 23.