Acta Scientiarum
http://periodicos.uem.br/ojs
ISSN on-line: 1807-8664
Doi: 10.4025/actascitechnol.v44i1.60486

Integrating corpus-based and NLP approach to extract terminology and domain-oriented information: an example of US military corpus

Liang-Ching Chen[1,2], Kuei-Hu Chang[3,4*] and Shu-Ching Yang[2]

[1]Department of Foreign Languages, ROC Military Academy, Kaohsiung, Taiwan. [2]Institute of Education, National Sun Yat-Sen University, Kaohsiung, Taiwan. [3]Department of Management Sciences, ROC Military Academy, Kaohsiung, Taiwan. [4]Institute of Innovation and Circular Economy, Asia University, Taichung, Taiwan. *Author for correspondence. E-mail: evenken2002@yahoo.com.tw

ABSTRACT. Within the modern information, communication and technology (ICT), seeking high efficient and accurate corpus-based approaches to process natural language data (NLD) is critical. Traditional corpus-based approaches for processing corpus (i.e. the collected NLD) mainly focused on quantifying and ranking words for assisting human in extracting keywords. However, traditional corpus-based approaches cannot identify the meanings behind the words to properly extract terminologies nor their information. To address this issue, the main objective of this paper is to propose an integrated linguistic analysis approach that combines two corpus-based approaches and a rule-based natural language processing (NLP) approach to extract and identify terminologies and create the text database for extracting deeper domain-oriented information by using the terminologies as channels to retrieve core information from the target corpus. Military domain is an uncommon research field and often classified as confidential data, which caused little researches to focus on. Nevertheless, military information is vital to national security and should not be ignored. Hence, to verify the proposed approach in extracting terminologies and information of the terminologies, the researchers adopt the US Army field manual (FM) 8-10-6 as the target corpus and empirical case. Compared with AntConc 3.5.8 and Tongpoon-Patanasorn's hybrid approach, the results indicate that from the perspectives of terminology identification, texts database creation, domain knowledge extraction, only the proposed approach can handle all these issues.

Keywords: Information; communication and technology (ICT); corpus-based approach; natural language data (NLD); natural language processing (NLP); military.

Received on August 9, 2021.
Accepted on December 14, 2021.

Introduction

Within the modern information, communication and technology (ICT), a huge number of texts have been created electronically, and stored in computers, thus, seeking high efficient and accurate linguistic analytical approaches to process natural language data (NLD) is critical. In today's digitized world, this makes accessing such data or texts both more convenient and efficient, and researchers are always searching for ways to boost the efficiency of extracting domain-oriented keywords and information from such big NLD (i.e. corpora). The main objective of this paper is to integrate two corpus-based and a natural language processing (NLP) approach for extracting terminologies and domain-oriented information. Corpus-based approaches are widely adopted for processing big corpora, and one of its significant features is keyword identification. Keywords are recognized as the indications, focus, and core of knowledge of texts or certain corpora. Corpus software based on its statistical algorithm (e.g. likelihood test) to calculate and to compare token's frequency in the target corpus and the benchmark corpus. The calculated results called 'keyness' that determine tokens' degrees of domain relevance (e.g. Li, 2016; Gilmore & Millar, 2018). Traditionally, AntConc 3.5.8 (Anthony, 2019), a type of corpus software, has been applied in many corpus-based researches and efficiently conducted token categorization, which is suitable for processing corpora.

NLP is categorized into a sub-field of artificial intelligence (AI), combines the fields of information technology (IT) and linguistics, and focuses on how to make computers understand and generate NLD to enhance interactions between humans, or between humans and machines (Gatt & Krahmer, 2018; Ghalibaf,

Nazari, Gholian-Aval, & Tara, 2019). Liddy (2001) defined NLP as consisting of natural language understanding (NLU) and natural language generation (NLG). Currently, NLP techniques have been utilized in diverse fields such as search engines (Garcelon, Neuraz, Benoit, Salomon, & Burgun, 2017), machine translation (Costa-Jussà, 2018), linguistics research (Sullivan & Keith, 2019), information retrieval (IR) (Vijayarajan, Dinakaran, Tejaswin, & Lohani, 2016), automatic medical diagnosis systems (Wu et al., 2018), and so on. NLP algorithms are diverse, but can be roughly divided into rule-based and statistic-based algorithms. For decades, these two types of algorithms have been utilized in many applications (e.g. Chinea-Rios, Sanchis-Trilles, & Casacuberta, 2019; Chang, 2017; Gablasova, Brezina, & McEnery, 2017; Hayat & Khan, 2012; Chang, 2021; Lu et al., 2021; Pichler, Boreux, Klein, Schleuning, & Hartig, 2019; Pota, Marulli, Esposito, De Pietro, & Fujita, 2019; Qin & Wang, 2020; Wang, Pakhomov, Ryan, & Melton, 2015; Wen, Chung, Chang, & Li, 2021; Xu, Shen, Liu, Zhang, & Shen, 2017; Chang, Wen, & Chung, 2018; Zhang, Boons, & Batista-Navarro, 2019).

One of the biggest limitations of traditional corpus-based approach (Anthony, 2019) is that AntConc 3.5.8 cannot distinguish function words, content words, nor domain-oriented keywords during corpus processing. Tongpoon-Patanasorn (2018) hence proposed a semi-automated corpus-based approach to firstly generate keyword list through AntConc 3.5.8, then, established a checklist to rate keywords for determining and extracting finance-oriented technical words. Although keywords can be properly retrieved by Tongpoon-Patanasorn's (2018) approach, a tailor-made text database that organized by language parser as a knowledge base for IR cannot be found from the traditional corpus-based approaches. To handle this issue, Chen (2013) based on Chomsky's (1957) linguistic theory, permutation, addition, deletion, substitution (PADS), to create an English parser as a rule-based NLP approach, which called computer reading system (CRS) for parsing English sentences by moving words or clusters to grammatical case frames automatically, and creating texts database by the machine's cumulated reading. However, the aforementioned approaches had their own pros and cons when processing corpus data. In this paper, the researchers hypothesize that a high efficiency linguistic analytical approach should be able to accurately extract domain-oriented keywords and use the keywords as the indexes (i.e. the channels) to retrieve domain-oriented information from its knowledge base. To effectively handle these issues, the main objective of this paper is to propose an integrated linguistic analysis approach that combines two corpus-based approaches and a rule-based NLP approach to extract and identify terminologies and create the text database for extracting deeper domain-oriented information by using the terminologies as channels to retrieve core information from the target corpus. To verify the proposed approach, the researchers adopted the U.S. Army field manual (FM) 8-10-2 (Headquarters Department of Defense Washington, DC, 2000) as a real-word military domain and the target corpus.

Methodology

Overview of the target corpus

US Army publications are open and released on the internet for military or non-military personnel to access for research or reference purposes. The U.S. Army FM 8-10-6, '*Medical Evacuation in a Theater of Operations: Tactics, Techniques, and Procedures*' (Headquarters Department of Defense, Washington, DC, 2000) provides professional knowledge and detailed explanations of medical evacuation procedures on operations for military personnel, and offers detailed tactics, techniques and procedures for the evacuation of casualties of sick, wounded and injured personnel. It is also one of the most important FMs at present US military.

This paper adopted the textual data of FM 8-10-6 as the target corpus and a real-world empirical example of military domain to verify the proposed approach.

The proposed integrated approach

Integrating different language analytical approaches can obtain the advantages of each approach and compensate each approach's deficiencies. To effectively process NLD for further extracting terminologies and the domain-oriented information, the proposed approach used in this paper integrates two corpus-based approaches and a rule-based NLP technique, which consists of seven steps, as shown in Figure 1; detailed descriptions of these steps are given below.

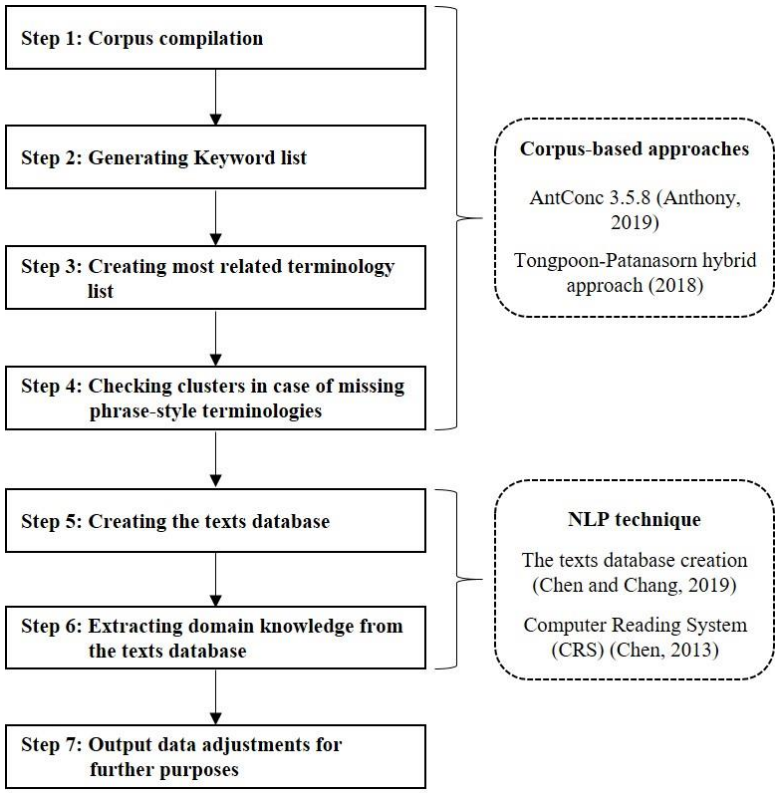


Figure 1. The flowchart of the proposed approach.

Step 1: Corpus compilation

The target corpus must be collected and transformed into .txt formats (both ANSI and UTF-8 types) for the corpus software to generate the keyword list and clusters of keywords, and for the NLP system to parse and create the texts database.

Step 2: Generating keyword list

To determine keywords, Dunning (1993) proposed the log-likelihood method to calculate a token's keyness value, which becomes an important algorithm in corpus software (see Definition 1).

Definition 1 (Dunning, 1993). Assume that two random variables, X_1 and X_2, follow the binomial distributions $B(N_1, P_1)$ and $B(N_2, P_2)$; P_1 and P_2 are a single trial's success probability, and N_1 and N_2 represent the number of successes that can occur anywhere, respectively. The likelihood ratio can be defined as:

$$-2\log\lambda = 2[\log L(P_1, K_1, N_1) + \log L(P_2, K_2, N_2) - \log L(P, K_1, N_1) - \log L(P, K_2, N_2)]$$

where

$$L(P, K, N) = P^K(1-P)^{N-K}$$

$$P_1 = \frac{K_1}{N_1},\ P_2 = \frac{K_2}{N_2},\ \text{and}\ P = \frac{K_1+K_2}{N_1+N_2}$$

AntConc 3.5.8 (Anthony, 2019), a corpus software, also adopted Dunning's (1993) log-likelihood method to calculate and determine the domain-oriented keywords by ranking tokens based on their keyness value.

In this step, the researchers extracted the domain-oriented keywords by generating keyword list through AntConc 3.5.8.

Step 3: Creating the most related terminology list

Once the keyword list is generated, based on Tongpoon-Patanasorn's (2018) checking list to rate each keyword's content-oriented levels and to eliminate words that are function words, meaningless words, unrelated to the topic, etc. The most related terminology list only embrace tokens that are rated as level three.

Step 4: Checking clusters in case of missing phrase-style terminologies

To avoid missing potential terminologies, based on the most related terminology list, check each term's clusters to find phrase-style terminologies through the corpus software, and the corpus software will automatically create a cluster list of a terminology.

Step 5: Creating the texts database

CRS (Chen, 2013), a NLP system, can mimic human readers' decision-making procedures and cognitive processes in reading. Its features include: (1) rule-based algorithm, (2) capable of sentence boundary detection, (3) capable of sentences segmentation, (4) capable of identify the event relations of each sentence, and (5) capable of parts of speech (POS) tagging. During this step, CRS (Chen, 2013) is used as an English parser to parse the target corpus. Figure 2 shows the text processing flowchart to illustrate how CRS (Chen, 2013) works. The target corpus is segmented from paragraphs into sentences, and from sentences into clauses, phrases and tokens. The small elements (i.e. tokens and clusters) will be distributed into the specific columns of the chart based on English linguistic rules (i.e. case frame and case grammar). All detailed output data can be accumulated to become the texts database (Chen & Chang, 2019).

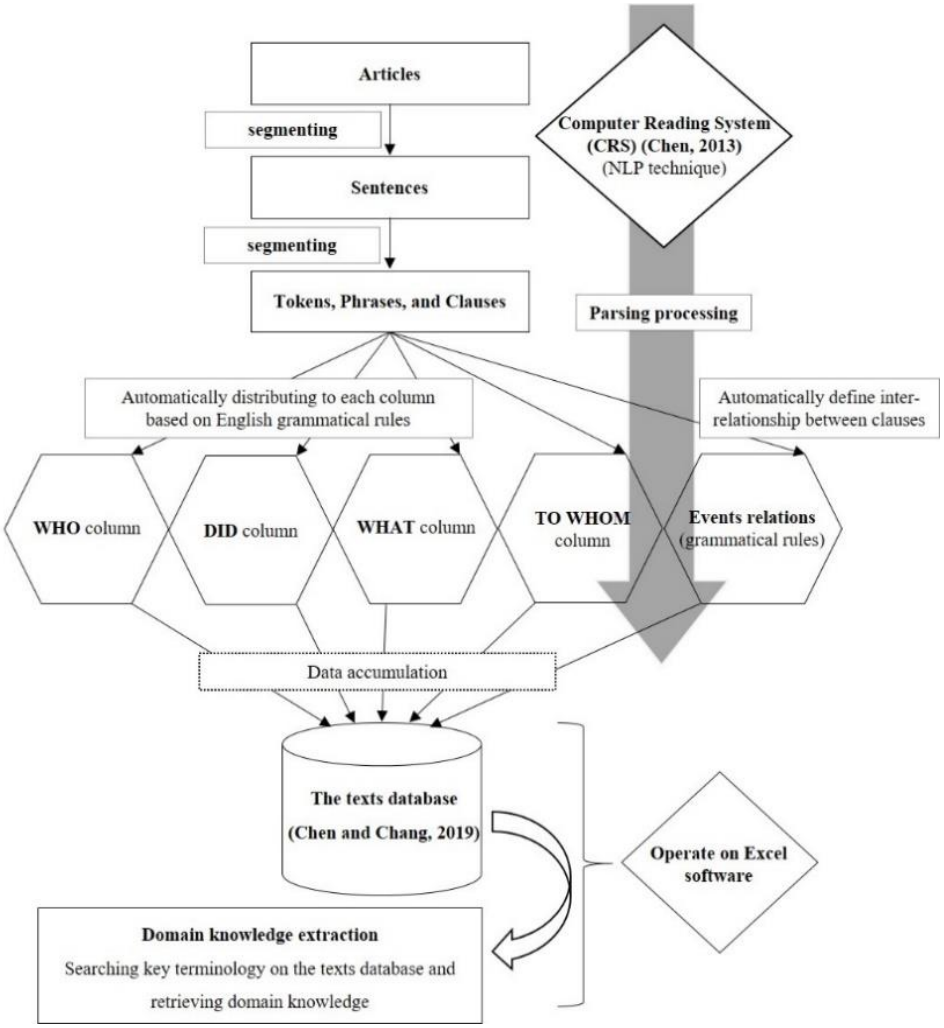


Figure 2. The flow chart of CRS parsing processing and the texts database creation.

Step 6: Extracting domain knowledge from the texts database

To explore further and detailed knowledge, the texts database (Chen & Chang, 2019) provides the service. The most related terminologies play as the indexes (i.e. channels) for extracting core information from the texts database.

Step 7: Output data adjustments for further purposes

The proposed integrated approach not only can be used for processing English NLD of any domain, but also its results can be tailored to suit for different purposes such as knowledge extraction, genre analysis, comparative analysis of terminology, etc.

Results and discussion

Results based on AntConc 3.5.8 (Anthony, 2019)

When the FM 8-10-6 was processed by AntConc 3.5.8 (Anthony, 2019), the keyword list (see Table 1) and illustrative examples of clusters of one most related terminology (see Table 2) indicate the important

linguistic evidences of the target corpus. However, AntConc 3.5.8 (Anthony, 2019) cannot process natural language with 100 percent correct to determine keywords, Table 1 indicated that some unnecessary data (i.e. function words) such as *the, must, be* showed on the keyword list, moreover, whether a keyword that determined by program is the core keyword of the domain is still unknown, hence, those still need to rely on the standardized manual tasks to optimize the data.

Table 1. An illustrative example of the keyword list (partial data).

Rank	Frequency	Keyness	Keywords	Rank	Frequency	Keyness	Keywords
1	627	6286.47	evacuation	26	206	543.64	must
2	640	4319.05	medical	27	86	532.59	corps
3	371	3530.35	litter	28	694	502.73	be
4	227	2175.43	casualty	29	48	499.83	aeromedical
5	296	1899.1	patient	30	98	495.4	unit
6	223	1884.32	ambulance	31	50	487.21	echelon
7	316	1831.69	patients	32	55	471.35	evacuated
8	238	1531.29	operations	33	66	454.71	rack
9	122	1270.59	CHS	34	82	442.92	vehicles
10	132	1230.18	ambulances	35	52	440.32	brigade
11	168	1191.31	combat	36	90	438.78	vehicle
12	266	1016.68	support	37	40	416.52	MRO
13	4192	930.14	the	38	117	409.1	available
14	127	795.95	personnel	39	43	403.54	turret
15	108	713.07	aircraft	40	506	400.17	or
16	187	700.23	air	41	65	380.01	hospitals
17	156	656.04	ground	42	74	377.44	theater
18	107	656.03	units	43	43	376.42	bearer
19	86	646.17	terrain	44	69	369.09	enemy
20	67	630.54	platoon	45	65	368.77	injured
21	63	610.62	bearers	46	68	368.68	rear
22	98	578.48	assets	47	95	361.65	movement
23	70	561.38	tactical	48	64	358.05	commander
24	108	556.9	equipment	49	44	356.81	battalion
25	75	552.26	casualties	50	91	355.89	required

Table 2. An illustrative example of noun phrase terminologies (adjective + nouns) (partial data).

Rank	Freq.	Range	Terms (NP.)	Rank	Freq.	Range	Terms (NP.)
1	204	9	medical evacuation	21	5	2	medical elements
2	51	8	medical personnel	22	4	2	medical operations
3	34	7	medical care	23	4	1	medical property
4	31	3	medical regulating	24	4	3	medical resupply
5	28	4	medical company	25	4	2	medical staff
6	22	5	medical treatment	26	4	4	medical support
7	20	3	medical group	27	4	2	medical unit
8	17	3	medical brigade	28	4	2	medical vehicles
9	17	7	medical supplies	29	3	1	medical commander
10	16	7	medical condition	30	3	3	medical intervention
11	14	2	medical platoon	31	3	1	medical training
12	13	5	medical units	32	2	2	medical assistant
13	12	4	medical companies	33	2	1	medical attendant
14	9	2	medical aidman	34	2	2	medical battalion
15	9	2	medical channels	35	2	2	medical command
16	7	3	medical equipment	36	2	1	medical department
17	7	2	medical intelligence	37	2	1	medical facilities
18	6	3	medical resources	38	2	1	medical facility
19	6	2	medical threat	39	2	2	medical items
20	5	2	medical capabilities	40	2	2	medical needs

Results based on the Tongpoon-Patanasorn hybrid approach (2018)

Although computer can successfully categorize tokens, it doesn't understand each word's importance level for people nor understand the meanings behind the words. Tongpoon-Patanasorn (2018) thus proposed a checklist as a criterion to filter terminologies, which divided terminologies into three different levels with the assistance of a check list. Level one considers words irrelevant to a specific domain, and are frequently used

in common purposes; level two considers words that may have certain definitions or explanations in a specific domain, and can be used in common or domain purposes; level three considers words that are significantly used in specialized purposes, and are strongly related to a specific domain. A keyword that identified as level three will be considered as a terminology (see Table 3).

Table 3. An illustrative example of the most related terminology list (partial data).

Rank	Freq.	Keyness	Terminology	Level	Rank	Freq.	Keyness	Terminology	Level
1	627	6286.47	evacuation	3	26	50	487.21	echelon	3
2	640	4319.05	medical	3	27	55	471.35	evacuated	3
3	227	2175.43	casualty	3	28	66	454.71	rack	3
4	296	1899.1	patient	3	29	52	440.32	brigade	3
5	223	1884.32	ambulance	3	30	40	416.52	MRO	3
6	316	1831.69	patients	3	31	43	403.54	turret	3
7	238	1531.29	operations	3	32	65	380.01	hospitals	3
8	122	1270.59	CHS	3	33	74	377.44	theater	3
9	132	1230.18	ambulances	3	34	43	376.42	bearer	3
10	168	1191.31	combat	3	35	69	369.09	enemy	3
11	266	1016.68	support	3	36	65	368.77	injured	3
12	108	713.07	aircraft	3	37	64	358.05	commander	3
13	187	700.23	air	3	38	44	356.81	battalion	3
14	156	656.04	ground	3	39	78	355.42	division	3
15	107	656.03	units	3	40	34	354.04	litters	3
16	86	646.17	terrain	3	41	42	336.47	evacuate	3
17	67	630.54	platoon	3	42	34	328.51	EAC	3
18	63	610.62	bearers	3	43	42	327.06	strap	3
19	98	578.48	assets	3	44	31	322.8	MTFs	3
20	70	561.38	tactical	3	45	75	318.96	mission	3
21	108	556.9	equipment	3	46	86	308.83	treatment	3
22	75	552.26	casualties	3	47	28	291.56	FSMC	3
23	86	532.59	corps	3	48	27	281.15	PCPS	3
24	48	499.83	aeromedical	3	49	31	274.62	medic	3
25	98	495.4	unit	3	50	38	262.62	FM	3

Results of the proposed integrated approach

The proposed approach integrates corpus-based approaches and a NLP technique to process military NLD. The following steps describe the completed results of the proposed approach.

Step 1: Corpus compilation

The target corpus contained 4,673 word types, 51,891 tokens, and its types token ratios (TTR) is 9% (see Table 4). Once the content was rearranged in .txt format, it had to be saved as ANSI and UTF-8 types, respectively. ANSI type was inputted to CRS (Chen, 2013), and UTF-8 type was inputted to AntConc 3.5.8 (Anthony, 2019).

Table 4. Lexical characters of the target corpus.

Corpus data	Word Types	Tokens	TTR
US Army FM 8-10-6	4,673	51,891	9%

Step 2: Generating Keyword list

The keyword list generator was run by AntConc 3.5.8 (Anthony, 2019), and the results are shown on Table 1.

Step 3: Creating the most related terminology list

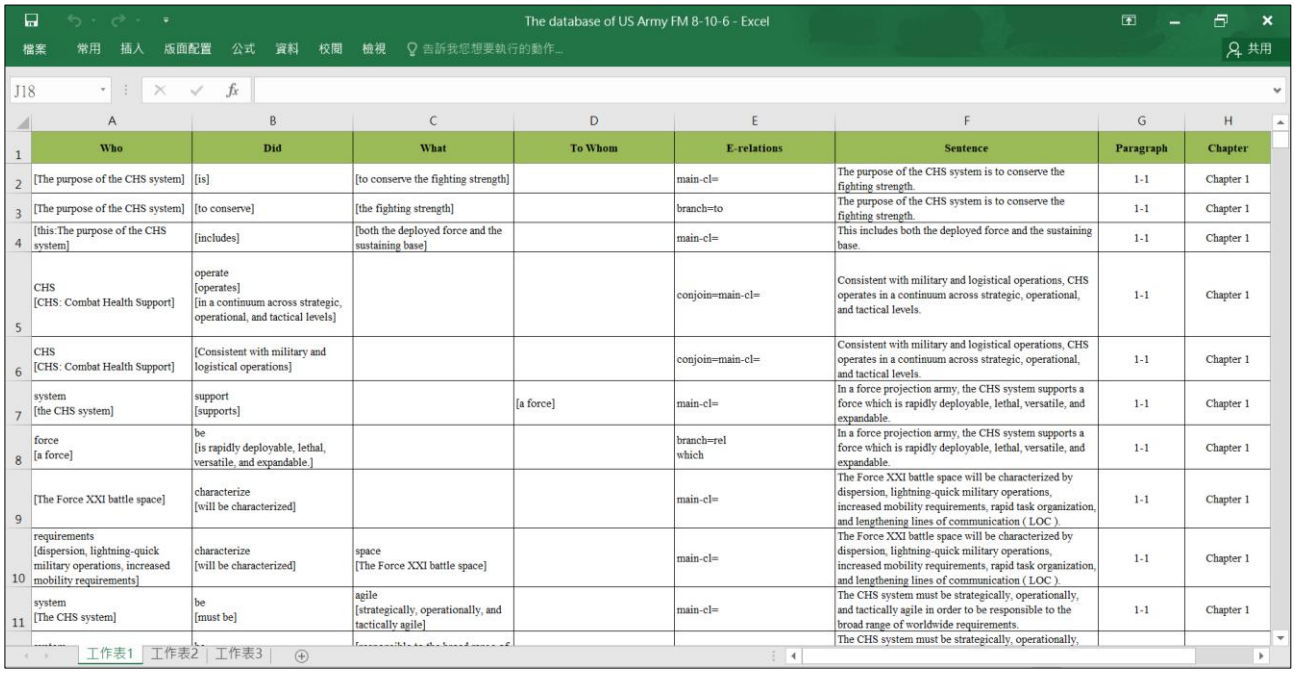
In this step, refining process is assisted by extracting the most related terminologies. Table 3 showed the partial data of the most related terminology list that labeled as level three by utilizing Tongpoon-Patanasorn hybrid approach (2018).

Step 4: Checking clusters in case of missing phrase-style terminologies

Once the most related terminology list is created, terminologies may exist in lexical bundle forms. Thus, multi-word combinations were checked by extracting phrase-style terminologies. For example, the researchers selected 'medical' on the most related terminology list, which is an adjective and have 640 frequency counts. Then clicking 'medical' in the corpus software to automatically generate a cluster list and emerged 79 derivative terminologies that are *medical* + nouns (see Table 2).

Step 5: Creating the texts database

Using CRS (Chen, 2013) to parse the target corpus and create the texts database (Chen & Chang, 2019). In the database, NLD of the FM was segmented into smaller elements (i.e. words, phrases, clauses) and those elements were distributed into different columns based on English linguistic rules automatically. The texts database provided the flexibility of searching key terminologies in different columns on Excel interface (see Figure 3). The results could base on different grammatical rules to yield various and detailed domain information.



	A	B	C	D	E	F	G	H
1	**Who**	**Did**	**What**	**To Whom**	**E-relations**	**Sentence**	**Paragraph**	**Chapter**
2	[The purpose of the CHS system]	[is]	[to conserve the fighting strength]		main-cl=	The purpose of the CHS system is to conserve the fighting strength.	1-1	Chapter 1
3	[The purpose of the CHS system]	[to conserve]	[the fighting strength]		branch=to	The purpose of the CHS system is to conserve the fighting strength.	1-1	Chapter 1
4	[this:The purpose of the CHS system]	[includes]	[both the deployed force and the sustaining base]		main-cl=	This includes both the deployed force and the sustaining base.	1-1	Chapter 1
5	CHS [CHS: Combat Health Support]	operate [operates] [in a continuum across strategic, operational, and tactical levels]			conjoin=main-cl=	Consistent with military and logistical operations, CHS operates in a continuum across strategic, operational, and tactical levels.	1-1	Chapter 1
6	CHS [CHS: Combat Health Support]	[Consistent with military and logistical operations]			conjoin=main-cl=	Consistent with military and logistical operations, CHS operates in a continuum across strategic, operational, and tactical levels.	1-1	Chapter 1
7	system [the CHS system]	support [supports]		[a force]	main-cl=	In a force projection army, the CHS system supports a force which is rapidly deployable, lethal, versatile, and expandable.	1-1	Chapter 1
8	force [a force]	be [is rapidly deployable, lethal, versatile, and expandable.]			branch=rel which	In a force projection army, the CHS system supports a force which is rapidly deployable, lethal, versatile, and expandable.	1-1	Chapter 1
9	[The Force XXI battle space]	characterize [will be characterized]			main-cl=	The Force XXI battle space will be characterized by dispersion, lightning-quick military operations, increased mobility requirements, rapid task organization, and lengthening lines of communication (LOC).	1-1	Chapter 1
10	requirements [dispersion, lightning-quick military operations, increased mobility requirements]	characterize [will be characterized]	space [The Force XXI battle space]		main-cl=	The Force XXI battle space will be characterized by dispersion, lightning-quick military operations, increased mobility requirements, rapid task organization, and lengthening lines of communication (LOC).	1-1	Chapter 1
11	system [The CHS system]	be [must be]	agile [strategically, operationally, and tactically agile]		main-cl=	The CHS system must be strategically, operationally, and tactically agile in order to be responsible to the broad range of worldwide requirements.	1-1	Chapter 1
						The CHS system must be strategically, operationally,		

Figure 3. A screenshot of the texts database.

Step 6: Extracting domain knowledge from the texts database

To dig deeper, information related to a terminology can be searched on the texts database. For example, the researchers chose 'casualty' which had 227 frequency counts, and typed it into the search bar of the 'Who' column. The results presented all sentences containing 'casualty' or clusters of 'casualty' as subjects (i.e. 'Who' column). The texts database gathered all information combined with 'casualty' from the target corpus. If 'casualty' is searched in the columns of 'Did', 'What', or 'To Whom', the results will be quite diverse because those represent different linguistic roles that a terminology plays.

Step 7: Output data adjustments for further purposes

The output data of the proposed approach can be converted to suit many purposes (e.g. developments of military training material, FM guide-reading materials, references information, etc.).

Discussion

This section uses NLD of U.S. Army FM 8-10-6 as the target corpus to discuss the differences in terminology identification, texts database creation and domain knowledge extraction between the AntConc 3.5.8 (Anthony, 2019), Tongpoon-Patanasorn's hybrid approach (2018) and the proposed approach (see Table 5). First, terminology identification was compared. Terminology extraction relies on manual identification. The AntConc 3.5.8 (Anthony, 2019) based on its statistical algorithm to calculate tokens' keyness and generate the keyword list which is considered as emerging domain-oriented words. However, keyword list still contained words for general purposes, function words, meaningless words and so on. Thus, it needed to be refined manually. The proposed method and Tongpoon-Patanasorn's (2018) hybrid approach utilized the same checking list to extract professional terminologies. Therefore, both approaches are able to identify terminologies from specific domains base on refining the results of the corpus software.

Second, the texts database creation was compared. Neither the AntConc 3.5.8 (Anthony, 2019) nor Tongpoon-Patanasorn's hybrid approach are designed for creating specific texts databases for specialized

usages. Those approaches basically focus on analysis of linguistic patterns. Nevertheless, the proposed method is designed for creating a specific texts database by a NLP technique, and inherits those approaches' advantages of corpus-based approaches, which indicated that it not only can capture most related military terminologies, but also can retrieve military domain knowledge from the texts database.

Finally, domain knowledge extraction was compared. Corpus-based approaches (Anthony, 2019; Tongpoon-Patanasorn, 2018) were limited to the analytical program's platforms, and it was difficult to retrieve the desired information flexibly. The proposed approach utilized CRS (Chen, 2013) to segment texts and create the texts database (Chen & Chang, 2019). When the texts database was estabilished on Excel software, it provided a highly flexible interface to extract domain knowledge from the target corpus.

Table 5. Comparisons of three methods.

Methods	Linguistic information retrieval		
	Terminology identification	Texts database creation	Domain knowledge extraction
AntConc 3.5.8 (Anthony, 2019)	No	No	Partial
Tongpoon-Patanasorn's hybrid approach (2018)	Yes	No	Partial
Proposed approach	Yes	Yes	Yes

Conclusion

NLD of military domain is difficult to process because it consists of complicated and uncommon terminologies, and its domain-oriented information is sometimes confidential. To handle this issue, the main objective of this paper is to propose an integrated approach that combines two corpus-based approaches and a rule-based NLP approach to extract and identify terminologies and create the text database for extracting deeper domain-oriented information by using the terminologies as channels to retrieve core information from the target corpus.

The significant contributions of the proposed approach are: (1) capable of generating the most related terminology list, (2) capable of identifying lexical bundles of terminologies, (3) capable of creating a tailor-made texts database, and (4) capable of retrieving detailed domain knowledge from the texts database. Feature one ensures the acquisition of core elements of the input data. Feature two unveils the domain-oriented terminologies by checking collocations. Feature three ensures the texts database is based on genuine material, and is also based on users' needs. Feature four extracts deeper domain information from the texts database. The proposed approach is designed to more efficiently conduct terminology extraction, information integration, and domain knowledge extraction.

In this paper, U.S. Army FM 8-10-6 was adopted as a real-world empirical case for verification; as demonstrated, the proposed integrated approach can be generally adopted to analyze NLD of different domains for extracting terminologies, establishing a tailor-made knowledge base, and adopting the terminologies as channels to retrieve domain-oriented information. Future researches can focus on improving corpus-based approaches, NLP tools, expanding the size of the database, optimizing the functions of the database, and so on.

Acknowledgements

The authors would like to thank the Ministry of Science and Technology, Taiwan, for financially supporting this research under Contract No. MOST 109-2410-H-145-002 and MOST 110-2410-H-145-001.

References

Anthony, L. (2019). AntConc (Version 3.5.8) [Computer Software]. Tokyo, JP: Waseda University.

Chen, S.-J. (2013). PADS restoration and its importance in reading comprehension and meaning representation. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation* [PACLIC], (p. 563–570). Retrieved from https://core.ac.uk/download/pdf/286947254.pdf

Chen, L.-C., & Chang, K.-H. (2019). Creating a novel type of information database of military terminology: an example of U.S. Army medical and casualty evacuation. *Basic & Clinical Pharmacology & Toxicology, 125*(S9), 8–8.

Chang, K.-H. (2017). A more general reliability allocation method using the hesitant fuzzy linguistic term set and minimal variance OWGA weights. *Applied Soft Computing, 56*, 589–596. DOI: http://doi.org/10.1016/j.asoc.2016.07.008

Chang, K.-H. (2021). A novel reliability calculation method under neutrosophic environments. *Annals of Operations Research, 2021*. DOI: http://doi.org/10.1007/s10479-020-03890-4

Chang, K.-H., Wen, T.-C., & Chung, H.-Y. (2018). Soft failure mode and effects analysis using the OWG operator and hesitant fuzzy linguistic term sets. *Journal of Intelligent & Fuzzy Systems, 34*(4), 2625–2639. DOI: http://doi.org/10.3233/JIFS-17594

Chinea-Rios, M., Sanchis-Trilles, G., & Casacuberta, F. (2019). Vector sentences representation for data selection in statistical machine translation. *Computer Speech and Language, 56*, 1–16. DOI: http://doi.org/10.1016/j.csl.2018.12.005

Chomsky, N. (1957). *Syntactic Structure*. The Hague, NL: Mouton.

Costa-Jussà, M. R. (2018). From feature to paradigm: deep learning in machine translation. *Journal of Artificial Intelligence Research, 61*, 947–974. DOI: http://doi.org/10.1613/jair.1.11198

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61-74.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: identifying, comparing, and interpreting the evidence. *Language Learning, 67*(S1), 155–179. DOI: http://doi.org/10.1111/lang.12225

Garcelon, N., Neuraz, A., Benoit, V., Salomon, R., & Burgun, A. (2017). Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. *Journal of the American Medical Informatics Association - JAMIA, 24*(3), 607–613. DOI: http://doi.org/10.1093/jamia/ocw144

Gatt, A., & Krahmer, E. (2018). Survey of the state of the art in natural language generation: core tasks, applications and evaluation. *Journal of Artificial Intelligence Research, 61*, 65-170. DOI: http://doi.org/10.1613/jair.5477

Ghalibaf, A. K., Nazari, E., Gholian-Aval, M., & Tara, M. (2019). Comprehensive overview of computer-based health information tailoring: a systematic scoping review. *BMJ Open, 9*(1), e021022. DOI: http://doi.org/10.1136/bmjopen-2017-021022

Gilmore, A., & Millar, N. (2018). The language of civil engineering research articles: a corpus-based approach. *English for Specific Purposes, 51*, 1–17. DOI: http://doi.org/10.1016/j.esp.2018.02.002

Hayat, M., & Khan, A. (2012). Discriminating outer membrane proteins with fuzzy K-nearest Neighbor algorithms based on the general form of chou's PseAAC. *Protein and Peptide Letters, 19*(4), 411–421. DOI: http://doi.org/10.2174/092986612799789387

Li, S. (2016). A corpus-based study of vague language in legislative texts: strategic use of vague terms. *English for Specific Purposes, 45*, 98–109. DOI: http://doi.org/10.1016/j.esp.2016.10.001

Liddy, E. D. (2001). Natural language processing. In *Encyclopedia of Library and Information Science* (2nd ed., p. 1-15). New York, NY: Marcel Decker.

Lu, Y., Yang, R., Jiang, X., Zhou, D., Yin, C., & Li, Z. (2021). MRE: A military relation extraction model based on BiGRU and multi-head attention. *Symmetry, 13*(9), 1742. DOI: http://doi.org/10.3390/sym13091742

Pichler, M., Boreux, V., Klein, A. M., Schleuning, M., & Hartig, F. (2019). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution, 11*, 281–293. DOI: http://doi.org/10.1111/2041-210X.13329

Pota, M., Marulli, F., Esposito, M., De Pietro, G., & Fujita, H. (2019). Multilingual POS tagging by a composite deep architecture based on character-level features and on-the-fly enriched word embeddings. *Knowledge-Based Systems, 164*, 309–323. DOI: http://doi.org/10.1016/j.knosys.2018.11.003

Qin, H., & Wang, Y. (2020). Enhancing named entity recognition from military news with bert. *Journal of Physics: Conference Series, 1453*, 012132. DOI: http://doi.org/10.1088/1742-6596/1453/1/012132

Sullivan, F. R., & Keith, P. K. (2019). Exploring the potential of natural language processing to support microgenetic analysis of collaborative learning discussions. *British Journal of Educational Technology, 50*(6), 3047–3063. DOI: http://doi.org/10.1111/bjet.12875

Tongpoon-Patanasorn, A. (2018). Developing a frequent technical words list for finance: A hybrid approach. *English for Specific Purposes, 51*, 45–54. DOI: http://doi.org/10.1016/j.esp.2018.03.002

Headquarters Department of the Army Washington, DC. (2000). Medical evacuation in a theater of operations: tactics, techniques, and procedures. Field Manual, (8-10-6).

Vijayarajan, V., Dinakaran, M., Tejaswin, P., & Lohani, M. (2016). A generic framework for ontology-based information retrieval and image retrieval in web data. *Human-Centric Computing and Information Sciences, 6*(18). DOI: http://doi.org/10.1186/s13673-016-0074-1

Wang, Y., Pakhomov, S., Ryan, J. O., & Melton, G. B. (2015). Domain adaption of parsing for operative notes. *Journal of Biomedical Informatics, 54*, 1–9. DOI: http://doi.org/10.1016/j.jbi.2015.01.016

Wen, T.-C., Chung, H.-Y., Chang, K.-H., & Li, Z.-S. (2021). A flexible risk assessment approach integrating subjective and objective weights under uncertainty. *Engineering Applications of Artificial Intelligence, 103*, 104310. DOI: http://doi.org/10.1016/j.engappai.2021.104310

Wu, J. T., Dernoncourt, F., Gehrmann, S., Tyler, P. D., Moseley, E. T., Carlson, E. T., ... Celi, L. A. (2018). Behind the scenes: a medical natural language processing project. *International Journal of Medical Informatics, 112*, 68–73. DOI: http://doi.org/10.1016/j.ijmedinf.2017.12.003

Xu, G., Shen, C., Liu, M., Zhang, F., & Shen, W. (2017). A user behavior prediction model based on parallel neural network and k-nearest neighbor algorithms. *Cluster Computing, 20*(2), 1703–1715. DOI: http://doi.org/10.1007/s10586-017-0749-z

Zhang, H., Boons, F., & Batista-Navarro, R. (2019). Whose story is it anyway? Automatic extraction of accounts from news articles. *Information Processing & Management, 56*(5), 1837–1848. DOI: http://doi.org/10.1016/j.ipm.2019.02.012