RESEARCH ON HEIDELTIME WITH VIETNAMESE LANGUAGE PROCESSING AND EXPERIMENTAL APPLICATION DEVELOPMENT AND EVALUATION

Dien Thi Hong Ha

University of Economics - Technology for Industries, Hanoi, Vietnam

ARTICLE INFORMATION

Journal: *Vinh University Journal of Science Natural Science, Engineering and Technology*
p-ISSN*: 3030-4563*
e-ISSN: *3030-4180*

Volume: 53
Issue: 4A
****Correspondence:***
dthha@uneti.edu.vn

Received: *01 October 2024*
Accepted: *22 November 2024*
Published: *20 December 2024*

Citation:
Dien Thi Hong Ha (2024). Research on Heideltime with Vietnamese language processing and experimental application development and evaluation. ***Vinh Uni. J. Sci.*** *Vol. 53 (4A), pp. 99-111 doi: 10.56824/vujs.2024a112a*

ABSTRACT

This paper presents software development for searching and extracting temporal information from text to help users access and understand content from electronic documents stored on organizational computer systems and websites using the HeidelTime tool. HeidelTime is a natural language processing tool customized to analyze temporal elements in Vietnamese contexts. The research methodology includes the following key steps: Surveying systems and user needs for time-based text search; analyzing and selecting natural language processing techniques, where HeidelTime is applied to identify and extract temporal information from Vietnamese text. The research results demonstrate that the time-based search software achieves high accuracy when deployed on the organization's document management system. It effectively supports time-based information retrieval and extraction, meeting users' practical needs. This study highlights the potential application of natural language processing technology in Vietnamese document management, contributing to improved storage and search efficiency within organizational information systems.

Keywords: Text extraction; natural language processing; information extraction; Vietnamese text; HeidelTime.

1. Introduction

In natural language processing (NLP), identifying and normalizing temporal expressions in text have garnered significant attention to enhance the effectiveness of information retrieval and management applications. One of the advanced tools in this domain is HeidelTime, developed by Strotgen and Gertz at Heidelberg University [1], [2]. HeidelTime is a multilingual temporal tagging system capable of extracting and normalizing time expressions in text according to the TimeML annotation format. This tool supports multiple languages, including Vietnamese, and can be applied to various types of documents such as news articles (NEWS), narratives (NARRATIVE), conversational language (COLLOQUIAL), and scientific texts (SCIENTIFIC). Notably, in the NEWS category, HeidelTime accurately

calculates temporal values based on the document's creation time; for instance, with a document created on March 26, 2024, the expression "*Three days later*" will be normalized to March 29, 2024. For the NARRATIVE and COLLOQUIAL genres, temporal expressions often take on a more generalized form due to the absence of specific time references.

In addition to HeidelTime, several other studies have developed tools for temporal recognition and processing [4], [5], [6], [7]. For example, the Stanford Temporal Tagger (SUTime) employs spelling rules and regular expressions to identify temporal expressions, achieving good results in English but lacking the capability to process Vietnamese. Other methods, such as DANTE, focus on temporal analysis based on machine learning models, which, while providing high performance, require extensive training datasets and are challenging to apply to low-resource languages like Vietnamese. Compared to other tools, a notable advantage of HeidelTime is its adjustable rule system tailored to each language, facilitating a more flexible and effective expansion of language support. However, a disadvantage of HeidelTime is its need for accuracy in handling temporal expressions with relative or ambiguous meanings in the NARRATIVE and COLLOQUIAL genres [8], [9], [10], [11].

This study focuses on exploring and applying HeidelTime in the context of Vietnamese within the NEWS genre, aiming to leverage the tool's strengths in processing clear and specific temporal elements.

2. Theoretical foundation

2.1. Processing workflow of HeidelTime

HeidelTime is constructed using Java programming and organized into distinct and independent packages to reflect its specific structure and processing workflow. The text processing procedure of HeidelTime consists of four main stages [3], [5], aimed at transforming the input documents into normalized documents with time annotations formatted according to TimeML.

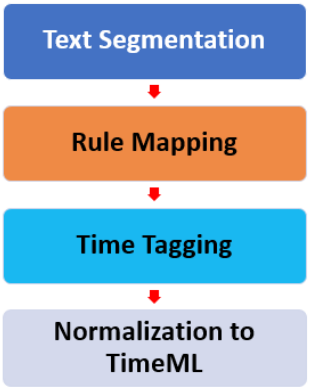

Figure 1: *Stages of document processing in HeidelTime*

As illustrated in Figure 1, the HeidelTime process comprises the following steps:

a. Text segmentation

The first step in the process is the segmentation and part-of-speech tagging of the text, which is particularly important for Vietnamese. In this step, the JVNTextPro tool

processes sentences and words, breaking them into grammatical units such as nouns, verbs, and adjectives. Due to the complexity of the language structure, JVNTextPro does not merely segment based on delimiters such as whitespace and punctuation; it also recognizes and assigns grammatical tags to the corresponding parts of speech. This crucial preprocessing stage in HeidelTime's analysis pipeline enables the system to understand and identify the grammatical structure of the text effectively.

b. Rule mapping

Following segmentation and part-of-speech tagging, HeidelTime maps the data to a resource of rules developed explicitly for Vietnamese. This rule set contains patterns, word values, and time recognition rules, supporting HeidelTime in determining and inferring the temporal values of related words. The resource is manually constructed to enhance accuracy in analyzing time-related elements characteristic of the Vietnamese language.

c. Time tagging

HeidelTime proceeds to tag temporal expressions in the text based on the mapping to the Vietnamese rule set. These tags not only contain the temporal values of the words but also identify their types and data formats. This stage is fundamental for normalizing time annotations in the output document.

d. Normalization to TimeML

Finally, HeidelTime normalizes the document into TimeML format, an annotation standard for temporal expressions. All words containing temporal elements are enclosed within <TIMEX3> tags, with attributes specifying their temporal values. An example from Figure 2 demonstrates how temporal expressions in the input text are formatted into TimeML, facilitating consistent and standardized storage and retrieval of time-related information.

2.2. Related theoretical background

This study is built upon the theoretical foundations of temporal annotation methods and natural language processing (NLP), mainly focusing on temporal tagging tools such as TimeML and the rules for normalizing time expressions in text. TimeML is an internationally recognized annotation standard for NLP, providing a unified structure for representing temporal elements within textual data. The JVNTextPro tool, designed for Vietnamese, has also been developed based on NLP principles, offering capabilities for analysis and part-of-speech tagging, which ensures effective processing of linguistically complex languages like Vietnamese.

Research on temporal recognition, such as SUTime and DANTE, has demonstrated the significance of rule-based and training data-based annotation methods. Compared to machine learning approaches that require extensive datasets, HeidelTime's rule-based method boasts advantages in flexibility and scalability for low-resource languages. These theories and tools serve as foundational elements for developing the HeidelTime application in the Vietnamese context [6], [8], contributing to refining and enhancing the accuracy of time information management and retrieval systems.

2.3. Development of rule resources for the Vietnamese language

HeidelTime operates based on rules and laws tailored to each supported language. This system enables HeidelTime to analyze and process text, producing output in the TimeML format. This study examines the rule system specific to the Vietnamese language. Within the HeidelTime software package, the "*resources*" directory contains language rule files, including Vietnamese ones (Figure 2).

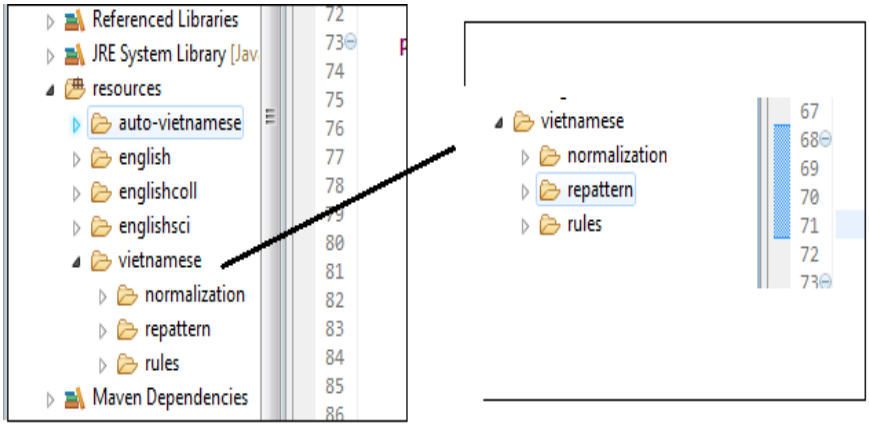

Figure 2: *Directory tree structure of HeidelTime*

In this study, we focus on the "Vietnamese" directory, which is similar to the language directories of HeidelTime, consisting of three subdirectories with specific functions as follows (Figure 3):

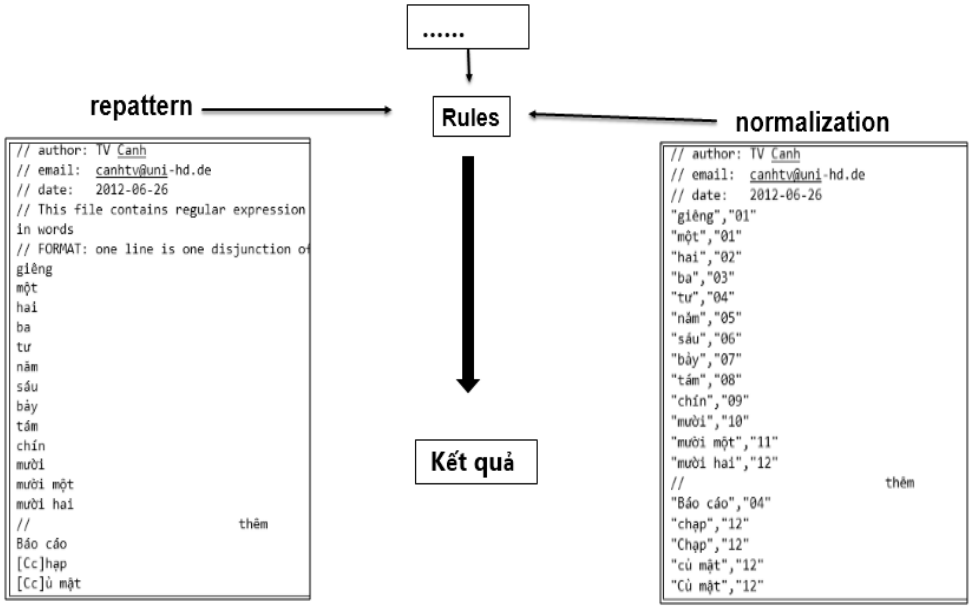

Figure 3: *Details of the three subdirectories of HeidelTime*

• **Repattern**: This directory contains commonly used time-related terms and allows access to all the rules in the "Rules" directory. These terms are organized in files

for easy use as a name catalogue rather than having to be fully listed. For example, the months from "tháng 1" (January) to "tháng 12" (December) are grouped into a single list file within this directory.

• **Normalization**: This directory contains the normalization values for the terms in the "Repattern" directory, which are compatible with the ISO standard time format. This normalization helps ensure consistency in the time information extracted from the text.

• **Rules**: This directory includes identifying, extracting, and normalizing expressions containing time elements in Vietnamese text.

To better understand the structure and functioning of these directories, we will examine each file in detail, such as the "reMonthInWord.txt" file located in the "Repattern" directory.

File ***reMonthInWord.txt*** in the ***repattern directory*** (Figure 4)

```
// author: TV Canh
// email:  canhtv@uni-hd.de
// date:   2012-06-26
// This file contains regular expression patterns for expressing months
in words
// FORMAT: one line is one disjunction of the pattern
giêng
một
hai
ba
tư
năm
sáu
bảy
tám
chín
mười
mười một
mười hai
//                              thêm
Báo cáo
[Cc]hạp
[Cc]ủ mật
```

Figure 4: *Details of the reMonthInWord.txt file*

File ***normMonthInWord.txt*** in the ***normalization directory*** (Figure 5)

```
// author: TV Canh
// email:  canhtv@uni-hd.de
// date:   2012-06-26
"giêng","01"
"một","01"
"hai","02"
"ba","03"
"tư","04"
"năm","05"
"sáu","06"
"bảy","07"
"tám","08"
"chín","09"
"mười","10"
"mười một","11"
"mười hai","12"
//                              thêm
"Báo cáo","04"
"chạp","12"
"Chạp","12"
"củ mật","12"
"Củ mật","12"
```

Figure 5: *Details of the normMonthInWord.txt file*

As we can see, the words found in the repattern file are also present in the normalization file, along with their corresponding values. Please note the writing style of ***[Cc]hạp*** and ***[Cc]ủ mật*** (Figure 6).

Repattern- reMonthInWord	Normalization- normMonthInWord
giêng một hai ba tư … Báo cáo [Cc]hạp [Cc]ủ mật	"giêng","01" "một","01" "hai","02" "ba","03" "tư","04" … "Báo cáo","04" "chạp","12" "Chạp","12" "củ mật","12" "Củ mật","12"

Figure 6: *Adding new words to both sides of the repattern and normalization directories*

As observed, the construction of the files containing the words we wish to define is simple; it is essential to ensure that additions are made to both the repattern and normalization directories. In the example shown in Figure 6, five new words have been added: "*Báo cáo, Chạp, chạp, Củ mật, củ mật***"**. This means that for the text containing the word "*Tháng Báo cáo***,"** HeidelTime will capture it, and the value assigned will be 04. Similarly, with "*tháng chạp, tháng củ mật"*.

But why could we capture it without even touching the rules directory? We added or modified entries in an existing file called *MonthInWord* (both in the *repattern* and *normalization* directories), where all the words in this file are defined by a typical formula within a single file in the rule's directory.

- Let us look at a definition formula for the *MonthInWord* file in the rules directory and analyze it to understand it better (Figure 7).

```
RULENAME="vn_date_r7h_lrec-BCADhint",
EXTRACTION="(tháng|Tháng) %reMonthInWord",
NORM_VALUE="%normMonthInWord(group(2))"
```

Figure 7: *Example definition formula for the* ***MonthInWord*** *file*

In the formula above, there are three main components:

+ RULENAME: The name of the rule.

+ EXTRACTION: HeidelTime will mark the words for temporal tagging when reading the text. These words come from the terms in the repattern directory, specifically from the reMonthInWord file. The components within EXTRACTION will be marked in order as groups 1, 2, etc. In this example:

- (tháng|Tháng) - is group (1).

- %reMonthInWord - is group (2).

+ NORM_VALUE: Extracts the value of the marked words from the normalization directory, which in this case corresponds to the normMonthInWord file.

With the example in Figure 7, HeidelTime will capture all words such as:

"Tháng giêng, tháng giêng, Tháng Báo cáo, tháng Báo cáo…"

Thus, we have understood the structure and the ability to add or modify the rule resources for the Vietnamese language, which is similarly applicable to all other languages.

3. Development and evaluation of the experimental program

After understanding the nature of HeidelTime and the rule resources for the Vietnamese language, we will build a software application that utilizes HeidelTime to summarize text [10], [11], [12]. Specifically, the application will read and process articles from a MySQL database (with articles sourced from the website "*https://vnexpress.net*") and summarize all of those articles. It will then extract all sentences containing time-related terms while ignoring others. The process is outlined as follows: Retrieving Articles from "*https://vnexpress.net*"): The "*https://vnexpress.net*") website provides an RSS file, which is written in XML format. This file contains <*item*> tags, each of which holds the following sub-tags: <*title*>, <*des*>, <*pubDate*>, and <*link*> to the detailed article. An RSS file typically contains an average of 25 <*item*> tags, corresponding to 25 articles published by *vnexpress.net*. We will use the JDOM library to facilitate the retrieval of content from the tags (or nodes) of the XML document. The required action is to connect to the Internet using the *java.net* package provided by Java and load all desired RSS files (specifically, we will load data from 2020 to the present). For each RSS, we will use JDOM to extract and store the necessary information in the database. To optimize database storage and simplify data retrieval, we will only load the RSS file and save four fields: <*title*>, <*des*>, <*pubDate*>, and <*link*> into the database. Subsequently, we will search the database based on the <*title*> and <*des*> fields. Using the <*link*> field, we will connect to the Internet again and load the entire ".*html*" page containing the detailed content of the article. Through the Jsoup library, we will filter the necessary content. Thus, we will obtain the article that needs to be analyzed using HeidelTime. An example of the RSS file and the tags containing the necessary content for the database is shown in Figure 8.

Components of the application:

• **Database:** A database to store the link, title, description, and publication date of articles from the *vnexpress.net* website. The database used here is MySQL. The articles from *vnexpress.net* include only those from news, education, and law (from 2020 to the present) sourced from the provided RSS file.

• **Lucene search engine:** Lucene is a lightweight yet powerful full-text search engine that enables rapid searching of relevant articles in the database. It collects data on each record, analyzes it, and indexes it, thus providing robust retrieval capabilities. While MySQL also offers querying capabilities, we choose to use Lucene due to the large volume of data and the need for fast search retrieval.

• **HeidelTime:** HeidelTime will process the documents, tag the time-related terms, and normalize them according to the TimeML standard. Since HeidelTime returns results in TimeML format, additional processing is required to make the results more

understandable, specifically ensuring that the output consists solely of sentences containing time elements and their corresponding time values.

• **JVNTextPro:** This is a preprocessing tool for the Vietnamese language and serves as the initial step of HeidelTime.

• **Rule resources for Vietnamese:** This component is packaged within HeidelTime.

• **JDOM:** This library is used to parse and extract content from XML-formatted documents' tags (or nodes), specifically the RSS files.

• **Jsoup:** Since the detailed article content is in .html format and does not conform to well-formed standards, extracting content by removing all tags using String methods or regular expressions can be pretty complex. Instead, we utilize Jsoup to facilitate the parsing of tags in the .html file and achieve the desired results.

• **MySQL Connector:** This component is used to connect to the MySQL database.

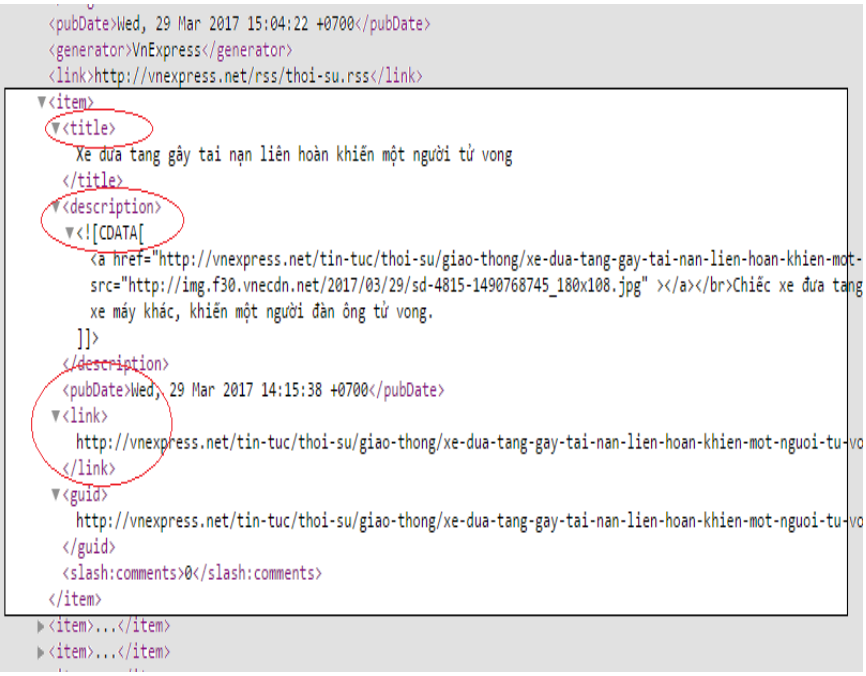

Figure 8: *RSS file with the necessary tags for the database*

Application functions:

• **Load data from vnexpress.net into the database:** Retrieve articles from the *vnexpress.net* RSS feed and store them in the database.

• **Search for text articles in the database:** Use the Lucene search engine to find articles that address the specific issues of interest.

• **Process text and documents:** Extract time-related terms and their corresponding values from user-input documents or those retrieved from the database. The resulting text can be formatted according to the TimeML standard or summarized to include only the sentences containing time-related terms.

• **Save/Export results:** Store or export the processed results to a file.

The structure of required classes is represented in Figure 9.

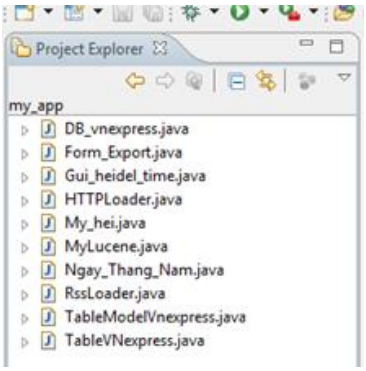

Figure 9: *Structure of the classes in the database*

Illustrative images (Figures 10, 11, 12, 13, 14, 15 and 16) showcasing the functionalities implemented in the software are as follows:

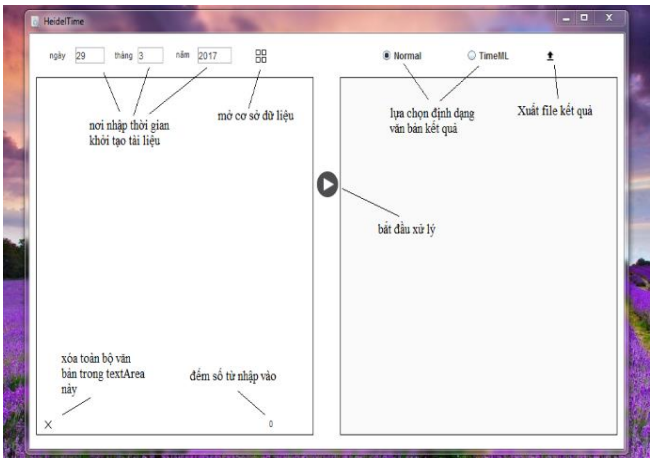

Figure 10: *Main interface*

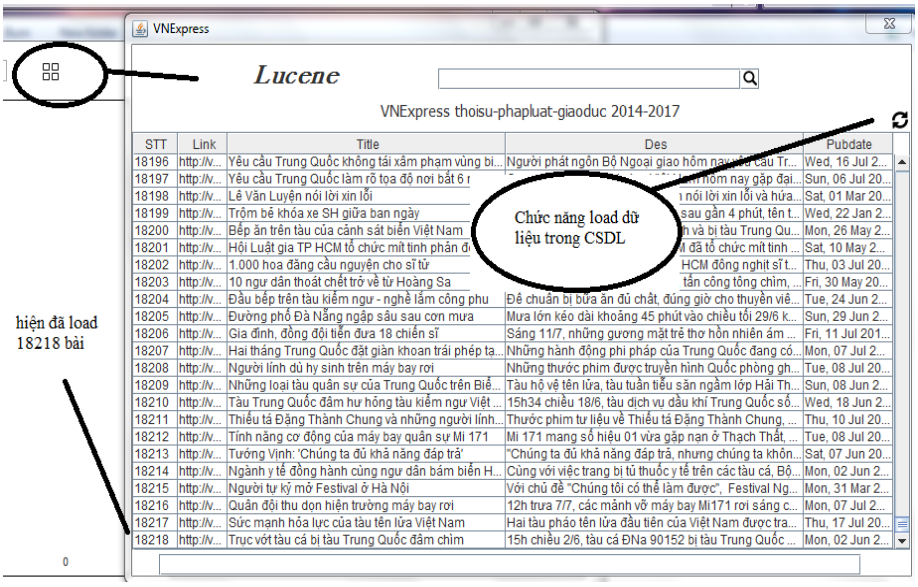

Figure 11: *When opening the database*

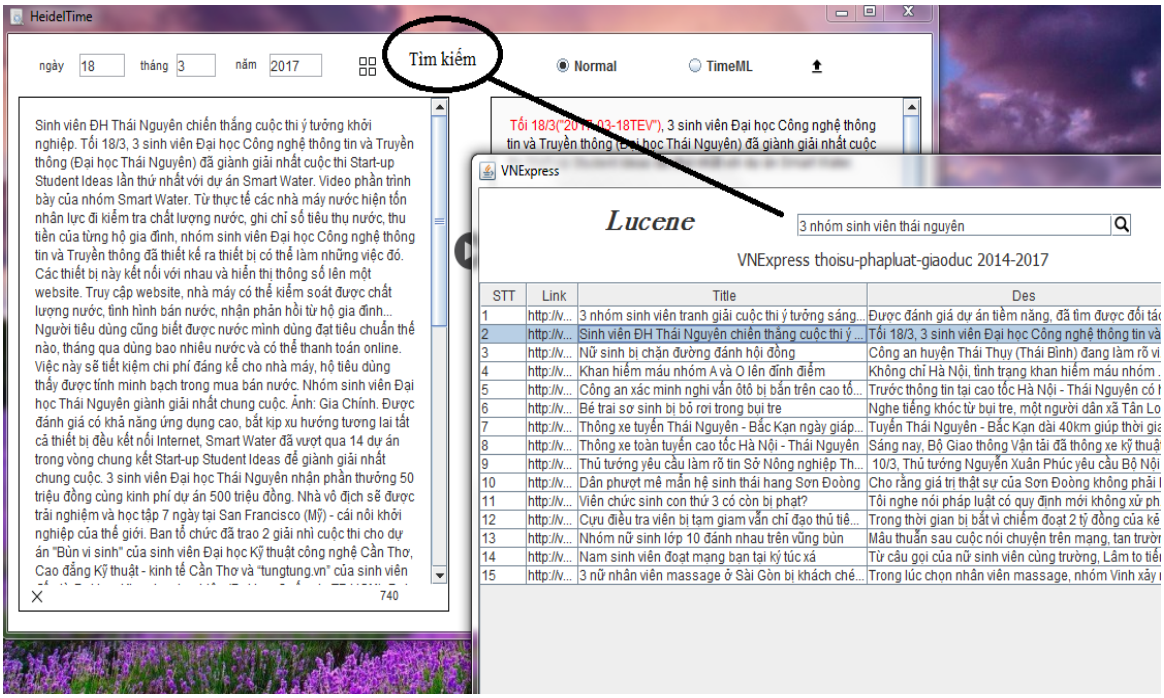

Figure 12: *Performing the search function with Lucene*

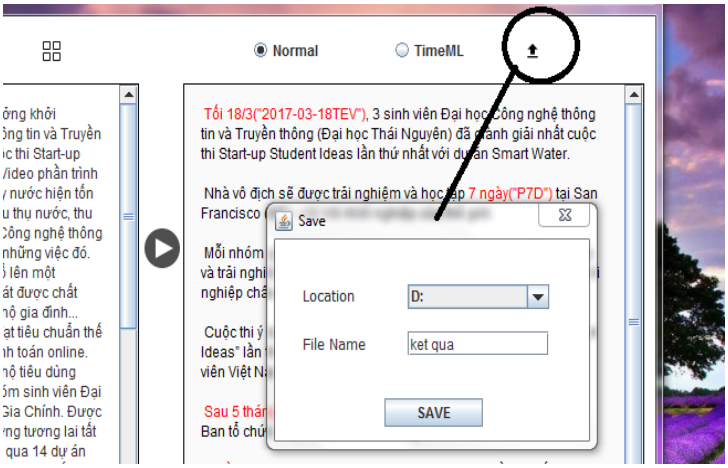

Figure 13: *Executing the function to export the result file*

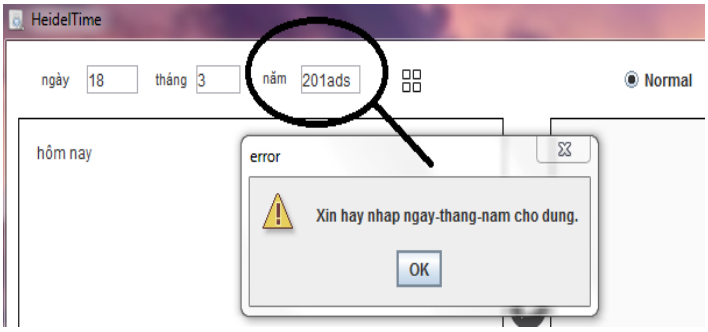

Figure 14: *Example of incorrect date entry (day-month-year)*

Figure 15: *Selecting the output format as Normal*

Figure 16: *Selecting the output format as TimeML (the standard format returned by HeidelTime)*

4. Conclusion

This study has developed a software application using the HeidelTime tool to search for and extract time-related information from articles on the *vnexpress.net* website, with data collected from 2020 to the present. The results indicate that the software can effectively process Vietnamese documents, accurately identifying and normalizing time-related expressions according to the TimeML format. The software facilitates quick access to important information for users and enhances document management efficiency within organizations. The application of the Lucene search engine significantly improves data retrieval capabilities, while the use of JVNTextPro ensures that the natural language processing aligns with the Vietnamese context. Future research directions will focus on improving text summarization capabilities and applying machine learning methods to develop models automatically and accurately extracting information. Additionally, expanding the application to other text types, such as scientific texts or financial reports, will be an important step toward enhancing the value of natural language processing technology across various fields.

REFERENCES

[1] Y. Philip, "XLTime: A Cross-Lingual Knowledge Transfer Framework for Temporal Expression Extraction," *Findings of the Association for Computational Linguistics: NAACL 2022*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2022, pp. 1931-1942. DOI: 10.18653/v1/2022.findings-naacl.148

[2] D. Wentao Ding, "A Pattern-Based Approach to Recognizing Time Expressions," In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, no. 1, pp. 6335-6342. DOI: 10.1609/aaai.v33i01.33016335

[3] J. Devlin, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," In *Proceedings of the 2019 Conference of the North*, Association for Computational Linguistics, Stroudsburg, PA, USA, 2019, pp. 4171-4186. DOI: 10.18653/v1/N19-1423

[4] P. Hausner, D. Aumiller, and M. Gertz, "Time-Centric Exploration of Court Documents," In *Proceedings of Text2Story - Third Workshop on Narrative Extraction From Texts Co-located with 42nd European Conference on Information Retrieval*, CEUR Workshop Proceedings, 2020, pp. 31-37.

[5] S. Jannik, "Adversarial Alignment of Multilingual Models for Extracting Temporal Expressions from Text," In *Proceedings of the 5th Workshop on Representation Learning for NLP*, Association for Computational Linguistics, Online, 2020, pp. 103-109. DOI: 10.18653/v1/2020.repl4nlp-1.14

[6] H. Sousa, "Tieval: An Evaluation Framework for Temporal Information Extraction Systems," In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, NY, USA, 2023, pp. 2871-2879. DOI: 10.1145/3539618.3591892

[7] L. Hui, S. Jannik, Z. Julian, and M. Gertz, "Chinese Temporal Tagging with HeidelTime," In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, vol. 2, Short Papers, pp. 133-137. DOI: 10.3115/v1/E14-4026

[8] S. Jannik and M. Gertz, "HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions," In *Proceedings of the 5th International Workshop on Semantic Evaluation*, ACL, 2010, pp. 321-324.

[9] H. Llorens, "TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2," In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval 2010)*, 2010, pp. 284-291.

[10] S. Jannik and M. Gertz, "Multilingual and Cross-domain Temporal Tagging," *Language Resources and Evaluation*, vol. 47, no. 2, pp. 269-298, 2013.

[11] S. Jannik, "HeidelTime: Tuning English and Developing Spanish Resources for TempEval-3," In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, 2013, pp. 15-19.

[12] N. UzZaman and J. Pustejovsky, "SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations," In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, 2013, pp. 1-9.

TÓM TẮT

NGHIÊN CỨU VỀ HEIDELTIME VỚI NGUỒN NGÔN NGỮ XỬ LÝ TIẾNG VIỆT VÀ XÂY DỰNG ỨNG DỤNG THỰC NGHIỆM, ĐÁNH GIÁ

Điền Thị Hồng Hà
Trường Đại học Kinh tế - Kỹ thuật công nghiệp, Hà Nội, Việt Nam
Ngày nhận bài 01/10/2024, ngày nhận đăng 22/11/2024

Bài báo này trình bày việc phát triển phần mềm tìm kiếm và trích xuất văn bản chứa yếu tố thời gian, nhằm hỗ trợ người dùng truy cập và hiểu nội dung từ các tài liệu điện tử lưu trữ trên hệ thống máy tính và website của tổ chức, sử dụng công cụ HeidelTime. HeidelTime là một công cụ xử lý ngôn ngữ tự nhiên được tùy chỉnh để phân tích yếu tố thời gian trong ngữ cảnh tiếng Việt. Phương pháp nghiên cứu bao gồm các bước chính: Khảo sát hệ thống và nhu cầu người dùng về tìm kiếm văn bản theo yếu tố thời gian; Phân tích và lựa chọn các kỹ thuật xử lý ngôn ngữ tự nhiên, trong đó HeidelTime được ứng dụng vào việc nhận diện và trích xuất thông tin thời gian từ văn bản tiếng Việt. Kết quả nghiên cứu cho thấy phần mềm tìm kiếm theo thời gian đạt độ chính xác cao khi triển khai trên hệ thống quản lý tài liệu của tổ chức. Phần mềm hỗ trợ tìm kiếm và trích xuất thông tin dựa trên yếu tố thời gian một cách hiệu quả, đáp ứng tốt nhu cầu thực tế của người dùng. Nghiên cứu khẳng định tiềm năng ứng dụng của công nghệ xử lý ngôn ngữ tự nhiên trong quản lý tài liệu tiếng Việt, góp phần nâng cao hiệu quả lưu trữ và tìm kiếm trong các hệ thống quản lý thông tin của các tổ chức.

Từ khóa: Trích xuất văn bản; xử lý ngôn ngữ tự nhiên; trích xuất thông tin; văn bản tiếng Việt; HeidelTime.