Efficient Language Identification for All-Language Internet News

Jian Tang
Laboratory of Language Engineering and Computing
Guangdong University of Foreign Studies
Guangzhou 510420, China
e-mail: 1557423887@qq.com

Xiaojiang Chen
Laboratory of Language Engineering and Computing
Guangdong University of Foreign Studies
Guangzhou 510420, China
e-mail: 774847467@qq.com

Wuying Liu*
Laboratory of Language Engineering and Computing
Guangdong University of Foreign Studies
Guangzhou 510420, China
e-mail: wyliu@gdufs.edu.cn

2021 International Conference on Asian Language Processing (IALP) | 978-1-6654-8311-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/IALP54817.2021.9675270

***Abstract*—The rapid development of language science and computing technology, especially the popularization of broadband Internet, has caused the explosion of all-language news to spread and communicate faster and faster. Among multi-modal news such as text, image, audio, and video, text news still accounts for the largest proportion of Internet news. In the face of more than 7,000 existing human languages, efficiently identifying the language of text news has become the most basic natural language processing technology, which can select accurate language processing methods for subsequent in-depth content processing and network public opinion analysis. Based on the idea of N-Gram, we designed and implemented a set of language identification methods suitable for all-language Internet news from two aspects: language training and language identification, and applied it to actual text news preprocessing. The language identification results of all-language Internet news show that our method has good recognition accuracy and efficiency.**

***Keywords*—Language Identification; Internet News; N-Gram; All-Language**

I. Introduction

Language identification [1] is mainly to determine the language type of a text. The general principle is to train a language classification model from an existing corpus, and then use the language classification model to predict or determine the language type of a text. Language identification plays an important role in the field of natural language processing. For example, the premise of machine translation is to accurately identify the language type of the source language, and this task must be completed by language identification technology.

There have been many language identification methods up to the present. In the early research, there were many language identification methods based on linguistics [2]. For example, a list of stop words was proposed and classified according to the degree of overlap between the document and the list of different languages [3]. The classification of text language is realized by calculating the occurrence probability of trigrams and phrases in sentences [4]. Subsequently, there is an investigation using common character strings in a specific language to perform regular expression matching to achieve language classification [5]. Although this kind of method is simple and fast, it relies on linguistic features and has poor language transfer classification capabilities.

With the development of natural language processing, language identification methods based on statistical model have become the most widely used methods in research, such as combining N-Gram language model and naive Bayes classifier for language identification [6]. Some researchers have implemented Langid [7] language identification toolkit based on this method, which has higher recognition speed and accuracy than TextCat [8]. In addition, there are the space vector model based on N-Gram character feature weight [9] and the graph structure method based on N-Gram [10]. The latter effectively uses the information of the word itself and the information between words to improve the efficiency of language recognition on short text. Later, some researchers have improved this algorithm [11].

Although there are many methods for language identification, they are rarely applied to actual work. In addition, the existing language identification tools are different in language recognition diversity and recognized objects, etc., which cannot meet the needs of existing research work well. Therefore, aiming at the Internet multilingual text news, using the existing language identification technology, we designed a set of language identification method of all-language Internet news, and applied it to the text preprocessing work of language identification and mark of the text data of foreign mainstream news media. This method can efficiently identify and mark a large number of all-language Internet news data, solve the problem of mixed multilingual text data, and make a certain contribution to the research in the field of natural language processing such as network public opinion analysis.

II. Algorithm Idea

A. N-Gram Algorithm

N-Gram refers to a continuous sequence containing N minimum segmentation units in a given text or speech sequence. The minimum segmentation unit can be phonemes, syllables, letters, words or some basic pairs customized according to specific applications. N-Gram is actually the representation of the N-1 Markov language model, so the divided sequence retains the order information between characters and words to a certain extent. Assuming a list of random variables S_1， S_2， …，

* Corresponding Author

S_m , if the probability of any one of the random variables S_i is only related to the preceding N-1 variables s_{i-1} 、 s_{i-2}、 ...、 s_{i-n+1}, that is[12]:

$$P(S_i \mid S_{i-n+1} \ldots S_{i-1}) = P(S_i \mid S_1 \ldots S_{i-1})$$

It is called a Markov process of order N-1. The N-Gram model takes all consecutive and overlapping N words as a unit and assumes it as an N-1 order Markov process. The significance of this hypothesis is that the occurrence of the Nth word is only related to the first N-1 words, and not related to any other words. The probability of the entire sentence is the product of the occurrence probabilities of each word, and the probability of these individual words can be obtained by counting the number of simultaneous occurrences of N words in the corpus. N-Gram theory is mainly used in the research of information retrieval, such as retrieval preprocessing, indexing, language identification and other pilot work, including the field of speech and text analysis.

The basic idea based on the N-Gram algorithm is that the text content is operated in a sliding window of size N according to the byte stream to form a sequence of word fragments of length 1N, and the frequency of occurrence of each word is counted separately. Therefore, the calculation amount of N-Gram statistics in the text increases as the value of N increases. When using N-Gram to extract word units in training and test texts, the basic requirement is that the extracted N-Gram units can cover the semantic words in the document. Choose the random number n order as the experimental parameter, and hope to adjust it in subsequent experiments.

B. *Similarity Weight Algorithm*

The key step of the text language identification model is to refer to the N-Gram in each language corpus to calculate the weight of each test text N-Gram. That is, if the word unit extracted from the test text exists in a certain multilingual model, a weight value is given according to the index position of the word unit in the language model. Otherwise, a penalty value PUNISHMENT is given. The PUNISHMENT here is a global parameter that can be fine-tuned according to the specific running results.

For various text documents, their composition follows the following basic facts: words are formed by words, sentences are formed by phrases, paragraphs are formed by sentences, and paragraphs are formed by paragraphs. Of course, there are cases where single characters become words and words become sentences. In addition, for semantic expression, it is necessary to resort to appropriate punctuation. When the meaning of the word is not considered, the word can be treated the same as the punctuation mark. Based on the above facts, any text document can be regarded as a collection, and the elements of the collection can be words, words, sentences, punctuation marks, paragraphs, and so on. Suppose the text length is n, then the set should contain n elements of length 1, n-1 elements of length 2, ..., 1 element of length n, these elements are n-Gram, which is It is formed by sliding a sliding window of size n from the start position of the text to the end position. The set contains (n-m+1) elements of length m (1<=m<=n), but considering the characteristics of the set, the same elements need to be removed, so the upper limit of the number of elements of length m is (n-m+1). When the sliding window is smaller, the number of elements deviates from (n-m+1). When the sliding window is larger, it is closer to (n-m+1). Obviously, the upper limit of the set size is. Therefore, the set can be expressed as follows:

$$D_i = \{e_1^1, e_2^1, \ldots, e_{n1}^1, e_1^2, \ldots, e_{n2}^2, \ldots, e_1^m, \ldots, e_{nm}^m, \ldots, e_1^n\}$$

Among them, n1, n2,..., nm, etc. respectively represent the number of elements of length 1, 2,..., m, and that the e_k^m element is the kth element of all elements of length m. Without considering the length of the element, for the convenience of presentation, the above set can be expressed as:

$$D_i = \{e_1, e_2, \ldots, e_k\} \qquad (k \le \frac{n(n+1)}{2})$$

The similarity evaluation function of document D_i and document D_j is defined as[13]:

$$S(D_i, D_j) = \frac{\sum_{k=1}^{n} F(e_k)W(e_k)}{\sum_{k=1}^{n} W(e_k)}$$

That is, if the element e_k does not exist in D_i and D_j at the same time, this element does not contribute to the similarity. e_k is the weight evaluation function of the element. If the element is selected in a certain way, the value of the function is mainly contributed by the content of the independent variable e_k itself; if the element is randomly selected, the value of the function is mainly determined by the independent variable The length of e_k contributes to the contribution. From here, it can be seen that the weight evaluation function is very important. In actual application, it is only necessary to randomly select N-Gram among the compared documents, and then determine whether the n-Gram is also in the reference document set, and calculate the corresponding contribution, and finally the similarity of the two documents can be judged.

III. METHOD STRUCTURE

In this paper, according to the task requirements, we designed and implemented a set of language identification methods suitable of all-language Internet news, and its general structure is shown in Figure 1. The language identification method of all-language Internet news is divided into two modules: language identification and language training. In the language training module, it is divided into three parts: data collection, training new languages and add configuration files. Similarly, language identification is divided into three parts: data collection, language identification and data storage.

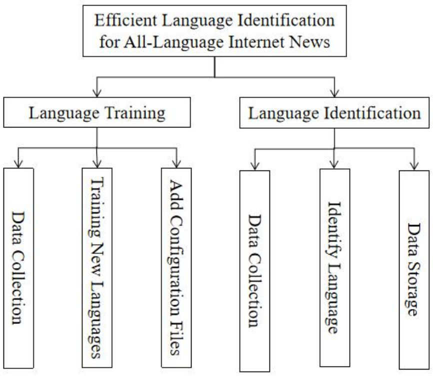


Figure 1. The General Structure of Language Identification for All-Language Internet News

A. Language Training Module

The structure of the language training module is shown in Figure 2. In this module, we can collect data according to the needs of language recognition, that is, collect the pure text data of the target language we want to recognize, and then store the text data in the local file. Taking the collected pure text data as the training text, we extract the N-Gram phrase by using the N-Gram algorithm, and get the language configuration file of the target language. Finally, we add the configuration file to the language configuration directory of the system, update the multi-language identification model, and then we can identify the language of the Internet news of the target language.

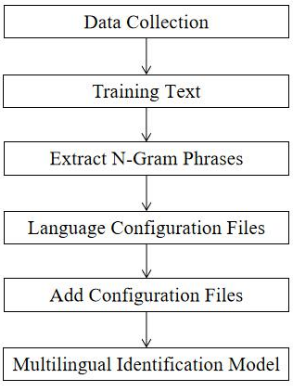


Figure 2. Structure Chart of Language Training

B. Language Identification Module

The structure of the language identification module is shown in Figure 3. The first thing we need to do is data collection, that is, the collection of all-language Internet news text data. After we get the text data, we transfer the data into the system, and then we can identify the language of the data. The system extracts N-Gram phrases from the incoming test text, and then calculates the similarity weight algorithm with the configuration file in the multi-language identification model to get the language identification results. Finally, we store the language identification results of the Internet news text data.

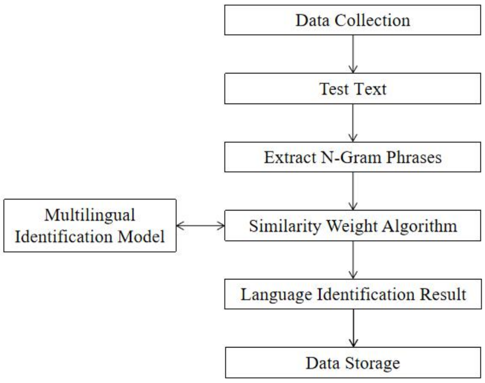

Data Collection
Test Text
Extract N-Gram Phrases
Multilingual Identification Model
Similarity Weight Algorithm
Language Identification Result
Data Storage

Figure 3. Structure Chart of Language Identification

In this paper, based on the idea of N-Gram, we designed and implemented a set of language identification methods. First, we extract N-Gram phrases from the training text to generate a language configuration file. Then we put the language configuration file into the configuration folder of the system to update the multi-language identification model. When we input the test text, we also perform N-Gram phrase extraction on it, and then calculate the similarity weight with the language configuration file in the multi-language identification model, judge its language and output the result. The model flow is shown in Figure 4.

The training text and the test text are subjected to the same N-Gram word unit extraction and word frequency statistics. The processed result of the training text is presented in the form of a text language model. After the test text is processed, the similarity weight algorithm will be used to calculate the similarity between it and each text language model, and the similarity will be used as the final judgment basis.

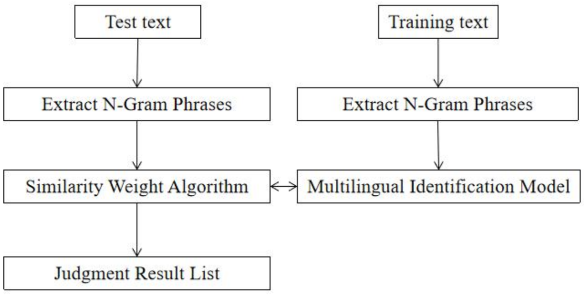
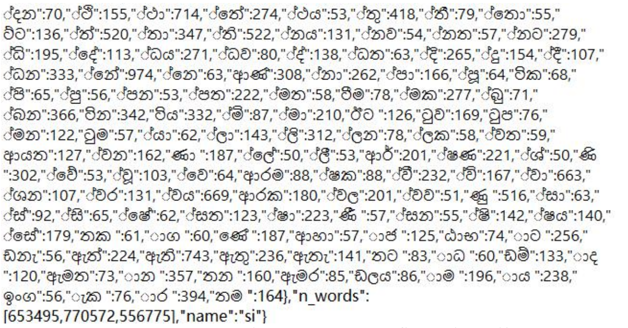


Figure 4. Model Flow Chart

IV. Implementation and Experiment

In this paper, we have implemented this method on Windows operating system. In the process of implementation, we need to install Java, Eclipse, Language detector, Jsonic, IO, JDK, slfj4-API and other installation packages. The language identification of all-language Internet news is divided into two modules. The operation of each module is independent and complementary. The following is the operation of each module and the data of language identification.

When we perform language identification on Internet news text data and find that we cannot identify the language of it, we can collect the target language text data by ourselves, and the method can be to obtain the target language data through major web translations. Then input the data into the language training module, and extract the N-Gram phrase from the data to obtain the language configuration file in “.json” format and add it, and then the multilingual identification model can be updated. As shown in Figure 5, this configuration file is a language configuration file for Sinhalese. Through the above operations, language identification can be performed on the Internet news text data in Sinhala, and the methods for adding other languages are the same as above.



```
්දන":70,"්ථි":155,"්ථා":714,"්තේ":274,"්ථය":53,"්තු":418,"්තී":79,"්තො":55,"
විට":136,"්ත්":520,"්තා":347,"්ති":522,"්තය":131,"්නව":54,"්නත":57,"්නට":279,"
්ධි":195,"්දේ":113,"්ධය":271,"්ධව":80,"්දි":138,"්ධත":63,"්දී":265,"්දු":154,"්දී":107,"
්ධන":333,"්නේ":974,"්නෙ":63,"ආණ්":308,"්නා":262,"්පා":166,"්පූ":64,"ටික":68,"
්පි":65,"්පු":56,"්පන":53,"්පත":222,"්මත":58,"ටීම":78,"්මක":277,"්බු":71,"
්බන":366,"ටින":342,"ටිය":332,"්මි":87,"්මා":210,"ඊට ":126,"වුව":169,"වුප":76,"
්මන":122,"වුම":57,"්යා":62,"්ලා":143,"්ලි":312,"්ලන":78,"්ලක":58,"්වත":59,"
ආයත":127,"්වන":162,"ණා ":187,"්ලේ":50,"්ලී":53,"ආර්":201,"්ෂණ":221,"්ශ්":50,"ණි
":302,"්වේ":53,"්වූ":103,"්වෙ":64,"ආරම":88,"්ෂක":88,"්වී":232,"්වි":167,"්වා":663,"
්ශන":107,"්වර":131,"්වය":669,"ආරක":180,"්වල":201,"්වව":51,"ණු ":516,"්සා":63,"
්ස්":92,"්සි":65,"්ෂේ":62,"්සත":123,"්ෂා":223,"ණී ":57,"්සන":55,"්ෂි":142,"්ෂය":140,"
්සේ":179,"තක ":61,"ාග ":60,"ණේ ":187,"ආහා":57,"ාජ ":125,"ඨාභ":74,"ාට ":256,"
ඩනැ":56,"ඇත්":224,"ඇති":743,"ඇතු":236,"ඇතැ":141,"තට ":83,"ාධ ":60,"ඩම්":133,"ාද
":120,"ඇමත":73,"ාන ":357,"තත ":160,"ඇමර":85,"ඩලය":86,"ාම ":196,"ාය ":238,"
ඉංග":56,"ැක ":76,"ාර ":394,"තම ":164},"n_words":
[653495,770572,556775],"name":"si"}
```

Figure 5. Sample Language Configuration File

The Internet text news used in this paper comes from major media news websites on the Internet and is obtained by crawling through crawler technology. The toolkits used in the crawling process mainly include Newspaper, General News Extractor, etc. The specific process is shown in the figure 6.

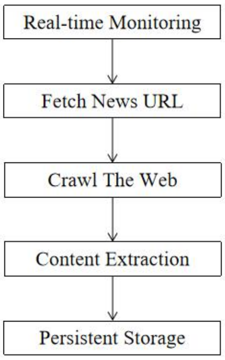
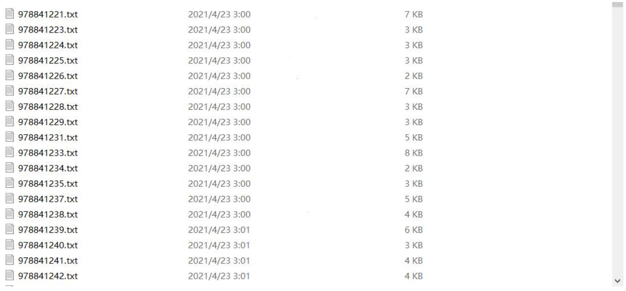


Figure 6. Data Collection Flow Chart

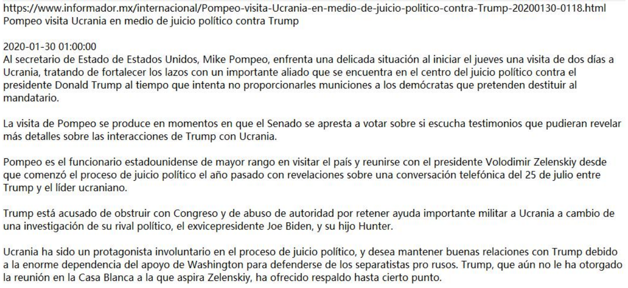
The first is to conduct real-time monitoring of major media news websites on the Internet, grab the URL of each news article in the news website, clean it, filter duplicate URLs, and store the news URL of the day. Then the parallel collection is carried out, and the news URL of the day is crawled through the Newspaper toolkit to obtain the content of each news page. Finally, the persistent storage operation is carried out, and the content of the crawled web page is extracted through the General News Extractor toolkit to obtain the time, author, title and content of each news. Each news is stored in a fixed format in a TXT document, and the daily news is stored in a separate folder.

The following shows the text data of Internet news. Taking the data of April 2021 as an example, figure 7 shows the text data of Internet news we collected on April 9, which is stored in a folder, and each TXT document represents a piece of news. Figure 8 shows an example of the text data of a piece of Internet news, including the URL, time, author and content of the news.



```
978841221.txt    2021/4/23 3:00    7 KB
978841223.txt    2021/4/23 3:00    3 KB
978841224.txt    2021/4/23 3:00    3 KB
978841225.txt    2021/4/23 3:00    3 KB
978841226.txt    2021/4/23 3:00    2 KB
978841227.txt    2021/4/23 3:00    7 KB
978841228.txt    2021/4/23 3:00    3 KB
978841229.txt    2021/4/23 3:00    3 KB
978841231.txt    2021/4/23 3:00    5 KB
978841233.txt    2021/4/23 3:00    8 KB
978841234.txt    2021/4/23 3:00    2 KB
978841235.txt    2021/4/23 3:00    3 KB
978841237.txt    2021/4/23 3:00    5 KB
978841238.txt    2021/4/23 3:00    4 KB
978841239.txt    2021/4/23 3:01    6 KB
978841240.txt    2021/4/23 3:01    3 KB
978841241.txt    2021/4/23 3:01    4 KB
978841242.txt    2021/4/23 3:01    4 KB
```

Figure 7. Daily Internet text News



https://www.informador.mx/internacional/Pompeo-visita-Ucrania-en-medio-de-juicio-politico-contra-Trump-20200130-0118.html
Pompeo visita Ucrania en medio de juicio político contra Trump

2020-01-30 01:00:00
Al secretario de Estado de Estados Unidos, Mike Pompeo, enfrenta una delicada situación al iniciar el jueves una visita de dos días a Ucrania, tratando de fortalecer los lazos con un importante aliado que se encuentra en el centro del juicio político contra el presidente Donald Trump al tiempo que intenta no proporcionarles municiones a los demócratas que pretenden destituir al mandatario.

La visita de Pompeo se produce en momentos en que el Senado se apresta a votar sobre si escucha testimonios que pudieran revelar más detalles sobre las interacciones de Trump con Ucrania.

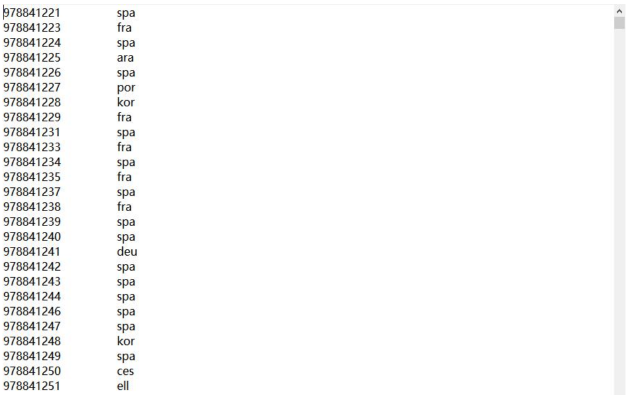
Pompeo es el funcionario estadounidense de mayor rango en visitar el país y reunirse con el presidente Volodimir Zelenskiy desde que comenzó el proceso de juicio político el año pasado con revelaciones sobre una conversación telefónica del 25 de julio entre Trump y el líder ucraniano.

Trump está acusado de obstruir con Congreso y de abuso de autoridad por retener ayuda importante militar a Ucrania a cambio de una investigación de su rival político, el exvicepresidente Joe Biden, y su hijo Hunter.

Ucrania ha sido un protagonista involuntario en el proceso de juicio político, y desea mantener buenas relaciones con Trump debido a la enorme dependencia del apoyo de Washington para defenderse de los separatistas pro rusos. Trump, que aún no le ha otorgado la reunión en la Casa Blanca a la que aspira Zelenskiy, ha ofrecido respaldo hasta cierto punto.

Figure 8. Internet text News Sample

After the test text is collected and stored locally, we only need to input the file address of the text file into the language identification module, which can identify the language of all the Internet text news of the day in order. By extracting n-gram phrases from the text data, we can calculate the similarity between it and the existing language configuration file. Find out the matching language, then output the result and store it persistently.As shown in Figure 9, which shows the language identification result of the Internet news on April 9, including the file name of the text news and the corresponding language.



```
978841221    spa
978841223    fra
978841224    spa
978841225    ara
978841226    spa
978841227    por
978841228    kor
978841229    fra
978841231    spa
978841233    fra
978841234    spa
978841235    fra
978841237    spa
978841238    fra
978841239    spa
978841240    spa
978841241    deu
978841242    spa
978841243    spa
978841244    spa
978841246    spa
978841247    spa
978841248    kor
978841249    spa
978841250    ces
978841251    ell
```

Figure 9. Language Identification Result Display Diagram

Table 1 makes statistics on the language identification results of Internet news on April 9. There were a total of 26,098 text news that day, and we accurately identified the language category of 25,727 news, and only 171 news failed to identify their language category. There are 61 kinds of language types for these news. From the statistical data, the method of language identification for all-language Internet news has good identification accuracy and efficiency.

TABLE I. LANGUAGE IDENTIFICATION RESULT

Language	Quantity	Language	Quantity	Language	Quantity	Language	Quantity	Language	Quantity
sin	1	**mar**	16	**dan**	75	**hun**	169	**ell**	776
bos	2	**isl**	21	**srp**	75	**lit**	170	**deu**	839
mlt	2	**mal**	25	**aze**	75	**ces**	187	**ind**	1101
afr	2	**kan**	28	**heb**	85	**nld**	204	**por**	1235
tha	2	**mon**	31	**cat**	91	**ben**	223	**ara**	1405
jpn	3	**eng**	37	**slv**	102	**hin**	237	**ita**	1543
swa	4	**nep**	39	**slk**	108	**bul**	324	**fra**	1995
kaz	4	**tel**	42	**nor**	113	**pol**	450	**rus**	4317
som	5	**fas**	47	**urd**	126	**hrv**	572	**spa**	5598
pan	9	**guj**	50	**hye**	130	**ukr**	610	**others**	171
msa	10	**mkd**	55	**fin**	131	**ron**	634		
tam	16	**est**	67	**swe**	140	**kor**	645		
glg	16	**lav**	67	**sqi**	154	**tur**	687		

V. CONCLUSION

In this paper, we designed and implemented a set of language identification methods suitable for all-language Internet news, which is used for language identification of text news. These text news come from foreign mainstream news media, and their language types are diversified and the amount of data is huge. This method has the following characteristics:

(1) All-language. For the text news on the Internet, its language types are diversified. In order to ensure the smooth progress of the follow-up research work on the text news, this method can identify all the languages of the Internet text news.

(2) Strong language scalability. This method has strong language expansion ability, we can easily add new languages according to the needs, and achieve the function of language identification for all text news .

(3) Efficient and simple. For the daily huge amount of Internet text news, this method can efficiently process the text data, the language identification operation is simple and convenient, and it has a good user experience.

ACKNOWLEDGEMENTS

The research is supported by the Social Science Foundation of Shanghai (No. 2019BYY028), the Humanity and Social Science Research Project of Ministry of Education of China (No. 20YJAZH069, No. 20YJC740062), and the Guangdong Basic and Applied Basic Research Foundation (No. 2018A030313672).

REFERENCES

[1] Y. Ikramu and W. Aishan, "Design and implementation of a highly concurrent language identification system for big data and short text," The Modern computer, pp. 7-13, 2020.

[2] B. Ding, "Research on language identification of social media short text based on N-Gram vector feature," Beijing University of Posts and Telecommunications, 2020.

[3] S. Johnson, "Solving the problem of language recognition," Technical report. School of Computer Studies, 1993.

[4] G. Grefenstette, "Comparing two language identification schemes," in Proc. of the 3rd International Conference on Statistical Analysis of Textual Data (JADT-95), 1995.

[5] C. Y. Feng and H. Y. Huang, "Multilingual identification based on character layer markov model," The Computer Science, vol. 33, no. 1, 2006.

[6] T. Vatanen, J. J. Väyrynen, and S. Virpioja, "Language Identification of Short Text Segments with N-gram Models," in LREC, 2010: Citeseer.

[7] M. Lui and T. Baldwin, "langid. py: An off-the-shelf language identification tool," in Proceedings of the ACL 2012 system demonstrations, 2012, pp. 25-30.

[8] W. B. Cavnar and J. M. Trenkle, "N-gram-based text categorization," in Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval, 1994, vol. 161175: Citeseer.

[9] R. D. Brown, "Selecting and weighting n-grams to identify 1100 languages," in International Conference on Text, Speech and Dialogue, 2013, pp. 475-483: Springer.

[10] E. Tromp and M. Pechenizkiy, "Graph-based n-gram language identification on short texts," in Proc. 20th Machine Learning conference of Belgium and The Netherlands, 2011, pp. 27-34.

[11] J. Vogel and D. Tresner-Kirsch, "Robust language identification in short, noisy texts: Improvements to liga," in Proceedings of the 3rd international Workshop on Mining Ubiquitous and Social Environments, 2012, pp. 1-9.

[12] M. Schonlau, N. Guenther, and I. J. T. S. J. Sucholutsky, "Text mining with n-gram variables," vol. 17, no. 4, pp. 866-881, 2017.

[13] Y. Gao, H. Zhao, Q. Zhou, M. Qiu, and M. Liu, "An Improved News Recommendation Algorithm Based on Text Similarity," in 2020 3rd International Conference on Smart BlockChain (SmartBlock), 2020, pp. 132-136: IEEE.