mathematics

MDPI

Article

A Privacy-Preserving Multilingual Comparable Corpus Construction Method in Internet of Things

Yu Weng [1], Shumin Dong [2,*] and Chaomurilige [1,*]

[1] Key Laboratory of Ethnic Language Intelligent Analysis and Security Governance of MOE, Minzu University of China, Beijing 100081, China; wengyu@muc.edu.cn
[2] School of Chinese Ethnic Minority Languages and Literatures, Minzu University of China, Beijing 100081, China
* Correspondence: 20400161@muc.edu.cn (S.D.); chaomurilige@muc.edu.cn (C.)

Abstract: With the expansion of the Internet of Things (IoT) and artificial intelligence (AI) technologies, multilingual scenarios are gradually increasing, and applications based on multilingual resources are also on the rise. In this process, apart from the need for the construction of multilingual resources, privacy protection issues like data privacy leakage are increasingly highlighted. Comparable corpus is important in multilingual language information processing in IoT. However, the multilingual comparable corpus concerning privacy preserving is rare, so there is an urgent need to construct a multilingual corpus resource. This paper proposes a method for constructing a privacy-preserving multilingual comparable corpus, taking Chinese–Uighur–Tibetan IoT based news as an example, and mapping the different language texts to a unified language vector space to avoid sensitive information, then calculates the similarity between different language texts and serves as a comparability index to construct comparable relations. Through the decision-making mechanism of minimizing the impossibility, it can identify a comparable corpus pair of multilingual texts based on chapter size to realize the construction of a privacy-preserving Chinese–Uighur–Tibetan comparable corpus (*CUTCC*). Evaluation experiments demonstrate the effectiveness of our proposed provable method, which outperforms in accuracy rate by 77%, recall rate by 34% and *F* value by 47.17%. The *CUTCC* provides valuable privacy-preserving data resources support and language service for multilingual situations in IoT.

Keywords: privacy protection; multilingual comparable corpus; Internet of Things

MSC: 68T01

Citation: Weng, Y.; Dong, S.; Chaomurilige. A Privacy-Preserving Multilingual Comparable Corpus Construction Method in Internet of Things. *Mathematics* **2024**, *12*, 598. https://doi.org/10.3390/math12040598

Academic Editors: Jialing He, Zhi Fang and Chunhai Li

Received: 11 January 2024
Revised: 11 February 2024
Accepted: 13 February 2024
Published: 17 February 2024

1. Introduction

With the continuous development of artificial intelligence (AI) and the widespread adoption Internet of Things (IoT) devices, the explosive growth of data generated by them provide AI with a vast amount of real-world data [1]. AI technology supports the intelligent analysis and provides abundant opportunities for the practical applications by the use of corpus [2]. However, the more corpus data provided, the more privacy and security challenges might occur during data processing and sharing [3,4]. Corpus, a set of well-sampled and processed electronic texts, is the basic resource for studies of linguistics and computational, especially language engineering for applications and devices in the IoT [5]. With the ever-widening coverage of IoT-related devices, the volume of data continues to increase, while the danger to privacy security increases simultaneously [6]. Thus, while the range of data information sources has become more extensive and precise, the involved language communities have become more complex and are faced with the potential risk of privacy protection. IoT devices can perceive useful multilingual data, and this information can facilitate intelligent decision-making [7], signifying a close connection between IoT and natural language processing. Multilingual comparable corpus, a collection of corpus data

consisting of texts in multiple different source languages that have comparable relationships in terms of topics or events, plays a vital role in this process. In this paper, the multilingual corpus, which makes use of IoT-based data as corpus material source, we call multilingual comparable corps in IoT. The challenge not only lies in delivering efficient multilingual interaction services within these complex language communities, and in transforming IoT data into accessible multilingual information, but most importantly, in the privacy and security protection of multilingual comparable corpus data sharing.

The privacy-preserving challenges of data sharing among the corpus have gained heated attention in AI and IoT [8,9]. In order to solve the privacy and security problems, the construction method of comparable corpus in IoT should take more issues into consideration, like how to protect original data values and ensure a certain level of privacy before publicly releasing datasets [3]. Based on this attempt, the construction procedures should ensure the security of the data source and the provability of the method. A parallel corpus is a text pair (or pairs) set consisting of the source language text and its corresponding translated text that needs to be aligned [10]. Compared to the parallel corpus, source language and target language texts of the comparable corpus are not strictly translatable and aligned. The acquisition of a comparable corpus is flexible, and the means of corpus construction is relatively convenient, so that the scale and the applied field have rapidly expanded [11], while the risk of collecting and sharing have also increased. In addition, as an important supplement to the parallel corpus, the comparable corpus has gradually become one of the indispensable research contents. Up to now, the comparable corpus has been widely used in translation equivalent extraction [12], multilingual sentiment analysis [13], machine translation [14], cross-language information retrieval, parallel sentence alignment [15], etc. Also, the extensive use of comparable corpora can be seen in various aspects in the IoT, for example, smart industry [16,17], smart city [18] smart transportation, news media [19] and so on [20,21]. With the continuous popularization and enrichment of applications on smart terminals, the use of multiple languages is gradually increasing and has led to a significant increase in multilingual and cross-lingual situations in IoT. However, IoT texts are typically concise, characterized by sparse content [22] and individual information or some sensitive information involved. When a parallel multilingual corpus is not feasible for IoT terminal devices, a comparable multilingual corpus serves as a valuable resource [23]. At present, Chinese and English, Russian, Japanese, and Spanish are common and frequently used in bilingual comparable corpora, while comparable corpora involving three or more languages are rare, and low-resource multilingual comparable corpora are even rarer. Although these tasks can be accomplished by using a two-by-two comparison approach based on multiple bilingual comparable corpora, the uneven distribution of a multiple bilingual comparable corpus and the tediousness of two-by-two comparisons greatly limit the quality and timeliness of these tasks.

To address the challenge of the timely provision of multilingual interactive services and to ensure the privacy and security of IoT-based information sharing in such multilingual communities, the utilization of a provable method of constructing multilingual comparable corpus becomes necessary. However, obtaining low-resource parallel corpora, particularly those involving minority languages, remains challenging, making the construction of a multilingual parallel corpus a formidable task. For this reason, this study initiates the construction of multilingual comparable corpora, including low-resource languages, as a foundational step towards the future extraction of high-quality multilingual potential parallel corpora, while also considering data-sharing privacy and security and facilitating rapid language interaction in non-strict translation scenarios. Aiming at the above problems, this paper proposes a provable method for building and implementing a privacy-preserving multilingual comparable corpus, involving the following steps: (1) projecting different language news text into a unified language text vector space; (2) based on privacy-preserving, solving the comparability problem among three or more different language news texts in this unified vector space; (3) deciding on the comparability of the three or more news texts based on an optimization decision mechanism with minimal impossibility

rule. According to the previous steps, we have realized a sample of privacy-preserving multilingual comparable corpus.

The main contributions of this paper are summarized as follows.

- We propose a similarity-based method for constructing a privacy-preserving multilingual comparable corpus in IoT involving three or more languages. Currently, natural language processing tasks often heavily rely on large-scale parallel corpora, while resources for multilingual parallel corpora are limited. A comparable corpus serves as a rich resource that offers indispensable supplements to parallel corpus. Previous research has predominantly focused on constructing bilingual comparable corpora, with little attention given to constructing comparable corpora involving three or more languages, particularly for low-resource languages. We introduce a capable approach to address this challenge of constructing a multilingual comparable corpus.
- We propose a decision making mechanism for comparing the comparability relationships in multilingual comparable corpora. In the existing process of constructing comparable corpora, there is limited research on decision mechanisms for comparability relationships. Most of the existing research primarily focuses on calculating the comparability between bilingual texts. However, when dealing with multiple languages, determining the comparability relationships between texts of different languages through simple calculations becomes challenging. Therefore, our proposed comparability decision making mechanism effectively addresses the issue of selecting comparable corpus pairs that satisfy comparability relationships across multiple languages and texts.
- The constructed corpus provides a better resource for the convenience of language activities like multilingual language teaching, compilation of multilingual dictionaries, cross-lingual translation studies, and a solution for the privacy and security challenges in IoT applications. From the perspective of privacy protection, during pre-processing, this corpus retains only elements such as titles and content, thus partially avoiding the retention and leakage of related privacy information. Furthermore, during sharing and using the corpus, by using the format of comparable pairs, users obtain processed and usable corpora instead of the original ones, which to some extent protects sensitive and personal information in the source corpus data, ensuring the privacy of the source language corpus.

The structure of this paper is as follows: Section 2 analyses and discusses the existing related work on construction methods and applications of privacy-preserving multilingual comparable news corpora in IoT. Section 3 presents the proposed provable construction method of a privacy-preserving multilingual comparable corpus in IoT. Section 4 describes the experiments and evaluation results. Section 5 concludes the current studies and outlines our future research directions.

2. Related Work

The related research on privacy-preserving comparable corpora in IoT can be divided into two main aspects: (1) research on key technical methods for privacy-preserving comparable corpus construction and comparability calculation, and (2) research on privacy-preserving multilingual comparable corpus applications in different occasions. IoT is a concept that aroused scholars' interest in 2018 [1,2], and which integrated variations of computing devices with different components for seamless connectivity and data transfer, including wearable devices [22], smart home appliances [24], IoT forensics [25], and applications for news broadcast and smart news media [26,27]. Since data have been gained and transferred more easily in the background of IoT, privacy and security are increasingly vital.

2.1. Methods of Privacy-Preserving Multilingual Comparable Corpus Construction and Comparability Calculation

A comparable corpus, as one of the important data resources and contents of corpus research, plays a more important role in linguistic research, especially in the related pro-

cessing of low-resource languages for smart devices in multilingual communities. And, the privacy and security of these data sources should be ensured. The relevant research on provable construction methods of privacy-preserving comparable corpora have been carried out at home and abroad [8], with the following three main methods: that based on word frequency distribution [28], that based on feature distribution, and that based on cross-lingual retrieval; e.g., D. Langlos et al. [29] used a cross-page semantic feature approach to obtain Arabic, English and French data and constructed a trilingual corpus, but the capacity of the corpus was comparably small, with only 305 comparable corpus pairs, and they ignored the privacy-preserving issue. The web-based construction method is a basic resource of comparable corpora, which mainly includes news websites; e.g., Yuan Wei [30] built a Chinese–Russian comparable news corpus by acquiring news corpora through the Xinhua website. Some other scholars focused on Wikipedia's corpus construction and some professional comparable corpus constructions based on domain websites. Based on previous studies, due to the time-consuming and laborious basic work of corpus construction, there are few self-built comparable corpora, and most of the comparable corpora are derived from off-the-shelf open source data. The data of comparable corpora based on existing corpora has limitations in specific domains and specialization, especially in specific domains such as the translation of classical works, simultaneous interpretation, and language teaching. There are limited existing multilingual comparable corpora accessible as a data set for a training model for IoT applications.

To satisfy the need for the privacy-preserving of corpus data, there are some methods proposed by scholars. When collecting and using these corpus data, problems related to privacy and security may arise. In response to this, some scholars have proposed data desensitization, encryption, and user authorization agreements. Data desensitization mainly involves replacing, masking, and confusing sensitive information, as well as using methods such as differential privacy [31] to minimize the risk. In the process of privacy-preserving comparable corpus construction, the establishment of comparable relations is also an essential part; meanwhile, comparability can be defined as the degree of similarity to some extent. Many scholars believe that comparability should satisfy the similarity between two comparative documents in terms of text length, style, domain, and temporal distribution. For example, Tan et al. [32] propose a method to calculate the similarity of bilingual sentences between Chinese and Lao by taking textual features such as parts of speech and numerical co-occurrence into consideration. At present, the classification methods of calculating the text comparability of monolingual and multilingual corpora that are more recognized by most scholars include the string-based method, dictionary-based method, corpus-based method, knowledge-based method, and hybrid methods [7]. Among them, multilingual comparability calculation can be carried out by dictionary-based methods [33], usually measured by the Jaccard similarity index [34], minimum edit distance algorithm, Dice coefficient [35], and other calculation methods, whose core idea is to transform the text similarity problem into a problem of collection by calculating the matching degree of terms directly through dictionaries. In recent years, scholars have started to apply neural networks to the field of natural language processing and have proposed neural network language models, of which word vectors are one of the methods of the research process of neural network language models. Word vectors represent text, words, sentences, paragraphs, and chapters as vectors, and they obtain the correlation between words by calculating the similarity between word vectors [25]. When bilingual or multilingual text is involved, translating source text into one unified language and converting it into vector space for similarity calculation has also become one of the methods for the comparability solving of comparable corpora [36].

Table 1 presents methods of privacy-preserving multilingual comparable corpus construction and comparability calculation. We develop our approach following the trends highlighted in the above review.

Table 1. Methods of Privacy-Preserving Multilingual Comparable Corpus Construction and Comparability Calculation.

Paper	Summary
[28]	Construct comparable corpus based on word frequency.
[29]	Construct trilingual comparable corpus based on cross-lingual retrieval, but ignore low resource languages.
[30]	Web-based construction method is employed to construct Chinese–Russian new comparable corpus.
[32,33]	Calculate comparability by taking textual features and dictionary-based methods.
[34,35]	Compute Jaccard similarity, minimum edit distance, and Dice coefficient as comparability index
[25,36]	Calculate the similarity between word vector.
Our study	We employ a news web-based construction method to gain source material and calculate comparability among multilingual news texts in a unified victor space.

2.2. Applications of Privacy-Preserving Multilingual Comparable Corpus in IoT

Many previous studies have found the close relationship between IoT data sharing and data privacy protection. Currently, the focus on data privacy protection is mainly in specific areas, such as the data analysis of patient conditions in the medical field [37], industry information [4] and data protection in the railway transportation sector [29]. There is relatively less coverage of privacy protection concerning multilingual data based on IoT news. The comparable corpus has gradually become an irreplaceable part due to its simple access and wide application; meanwhile, the privacy and security of multilingual comparable corpora has become an unavoidable part among data sharing and applications. Its applications are mainly focused on the following aspects: linguistic research mainly including the analysis of the translation skills, and comparative linguistic research [38]. Regarding computational research, its application is always about machine translation [39] and smart applications or devices in the IoT. Specifically, with the help of the IoT, news media has changed a lot involving the integration of media with big data technologies. At the same time, the danger of lost private information and sensitive data in IoT devices of various types, along with the valuable information they provide, have become integral to news reporting, particularly in the coverage of sudden news events [19,23]. Currently, news generated through IoT information has become commonplace, encompassing environmental monitoring, public service, and investigative news. In the current practical news application in the IoT, various multilingual scenarios arise, including traffic information broadcasting, emergency event reporting, and applications related to natural disasters. For example, taking vehicular networks as an example, Chang and Pan [39] propose a real-time dynamic news pattern concerning a vehicle traffic dynamic information network model and the IoT news reflecting traffic conditions, which can report congestion and complex road conditions, enabling individuals to decide whether to queue on the road, while vehicles can plan their routes based on the complexity of the ground road situation. Also, in Wang's [40] study, they collected data from Chinese news organizations to analyze the impact of the IoT on news media, which improved the security, team collaboration, high-speed network access, and public accessibility.

In O'Shaughnessy and Lin's study [3], they propose research on privacy protection practice for data mining with multiple data sources, putting forward a framework for reconstructing data and putting it to use in data clustering projects. Leveraging the IoT, information collection can be conducted at any time, from anywhere, using various devices, and through any transmission channel, which shows the necessity of privacy-preservation [6].

Consequently, news content automatically or semi-automatically includes text, pictures, audio, and video, before being stored. During this process, a comparable corpus is useful for information processing. In spite of the above occasions, a comparable corpus also can be extended to a multilingual situation, which is very useful in many IoT applications, such as language education, automatic translation and transcription, and multilingual customer service [24,40] like subway broadcast [41,42] and so on; for example, Duan et al. [22] propose an attempt at smart education for professional English teaching and provide a new teaching environment by designing different data transmission channels to improve applicability at different learning stages with the help of the IoT. However, most of the existing applications of the IoT involve general languages like English or Chinese, while fewer involve low-resource languages, and the development of foundational resources for multilingualism remains limited.

Table 2 shows the existing research on applications of multilingual comparable corpora in the IoT, including the privacy-preserving aspects in the above review.

Table 2. Applications of Privacy-preserving Multilingual Comparable Corpus in IoT.

Paper	Summary
[3,6,22]	Privacy protection practice for data mining and construction corpus for multiple data clustering; multilingual comparable corpus used in language teaching; evaluation of privacy-oriented corpus by use of text anonymization.
[4,8,30,36]	Privacy protection for medical data, industry data and railway data, less coverage of multilingual comparable data in IoT.
[12,13,22,38–40,42]	Multilingual comparable corpus used in machine translation, sentiment analysis; IoT-based corpus used in smart education; IoT-based data used in news media;
[36,43–45]	These studies focus on bilingual (common languages) comparable corpus; however, they do not mention multilingual, especially low-resource language comparable corpora.
Our study	We focus on the construction of three or more language comparable corpora that also meet the needs of privacy protection, which can be used in related multilingual situations, including all above aspects.

When discussing the languages involved in a comparable corpus, the existing foreign comparable corpus research focuses on the construction of common languages, and the comparable corpus has been built between Chinese and Thai, Cambodian, Vietnamese, Burmese, Russian and other languages along the Belt and Road [43]. However, there have been limited studies on low-resource languages, including ethnic language corpora involving a comparable corpus. Existing research mainly focuses on the construction and application of a monolingual or parallel corpus [36,44]. The current presentation of the ethnolinguistic corpus is mostly phonetic [45], textual, and a few multi-modal datasets, and the granularity of the corpus is mostly at the syllable, phoneme, word, or sentence level, with less research on chapter-level granularity and involving the linguistic information processing level of the corpus or database, and there is a deficiency in the design of shared applications of corpus resources.

In the context of the ongoing advancement of the IoT and news, facilitating the efficient exchange of barrier-free information in different (including ethnic) languages and ensuring the privacy and security of information play a crucial role in both production and daily life. Given this significance, the construction and utilization of privacy-preserving multilingual comparable corpora hold immense importance. Thus, this study aims to propose a provable method for constructing a privacy-preserving multilingual comparable

corpus within the IoT context and takes Chinese, Uighur, and Tibetan news as examples to verify its availability.

3. The Construction Method of Privacy-Preserving Multilingual Comparable Corpus

This study aims to propose a provable construction method for a privacy-preserving multilingual comparable corpus. As we mentioned before, our method has applied a web-based construction approach to gain corpus material. Before we discuss the main steps of our method, we describe the former procedures for gathering multilingual raw material and pre-processing a raw corpus in Figure 1:

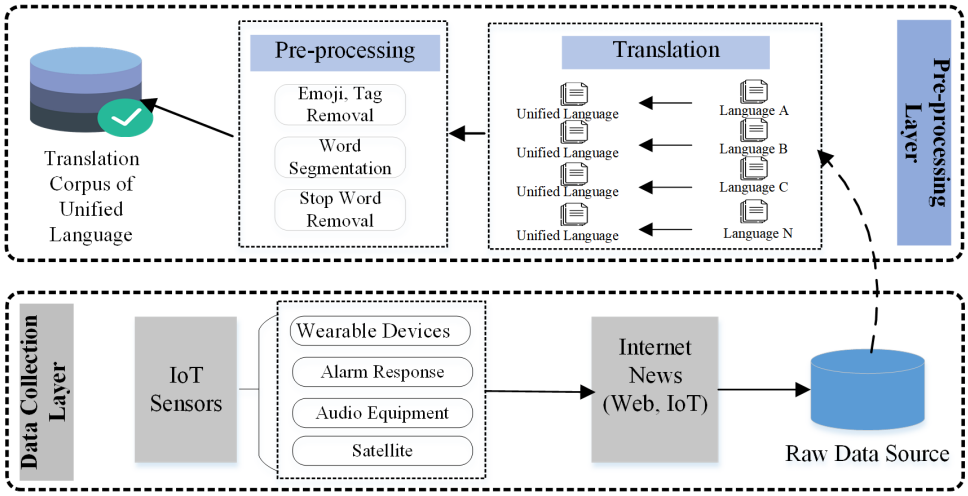

Figure 1. The pre-processing process of privacy-preserving multilingual comparable corpus.

In this part, we only choose three languages as a sample; in practical application, we can add more than three languages. Based on the former step, we have attained a unified language-described raw corpus. Our method consists of the following parts: (1) in order to efficiently calculate the comparability and ensure the privacy of the three different original language news data, we first embed the three or more different original language texts into the text vector space described in the unified language to form the corresponding embedded text vectors; (2) in the text vector space described in the unified language, we calculate the comparability between each pair of triplets with corresponding embedded text vectors; (3) based on the result of the comparability calculation in the procedure (2), we use a minimization principle based on impossibility to decide whether the triplet is a comparable pair; if they are comparable, they can be entered into the comparable corpus; otherwise, they are discarded. Each component of the construction method is described in detail below. Figure 2 shows all of the process and possible application situations in detail.

Suppose that $C = \{C_1, C_2, \ldots, C_K\}$, $U = \{U_1, U_2, \ldots, U_M\}$, $T = \{T_1, T_2, \ldots, T_L\}$ denotes the language $C(C)$, language $U(U)$ and language $T(T)$ raw news collections respectively, where $C_k(k \in [1, 2, \ldots, K])$ denotes the number kth document in the C raw news text collection, K denotes the number of documents; $U_m(m \in [1, 2, \ldots, M])$ denotes the number mth document in the U raw news text collection, M denotes the number of documents; $T_l(l \in [1, 2, \ldots, L])$ denotes the number lth document in the T raw news text collection,and L denotes the number of documents. Our task is to construct a multilingual comparable corpus of language C, U and $T(CUTCC)$, in other words, construct CUTCC = $\{(C_{ik}, U_{im}, T_{il})\}_N$, where C_k , U_m and T_l respectively denote the number Kth document in the raw C news text collection, the number mth document in the raw U news text collection, and the number lth document in the original T news text collection occurring in the CUTCC as the number i record, and they are comparable.

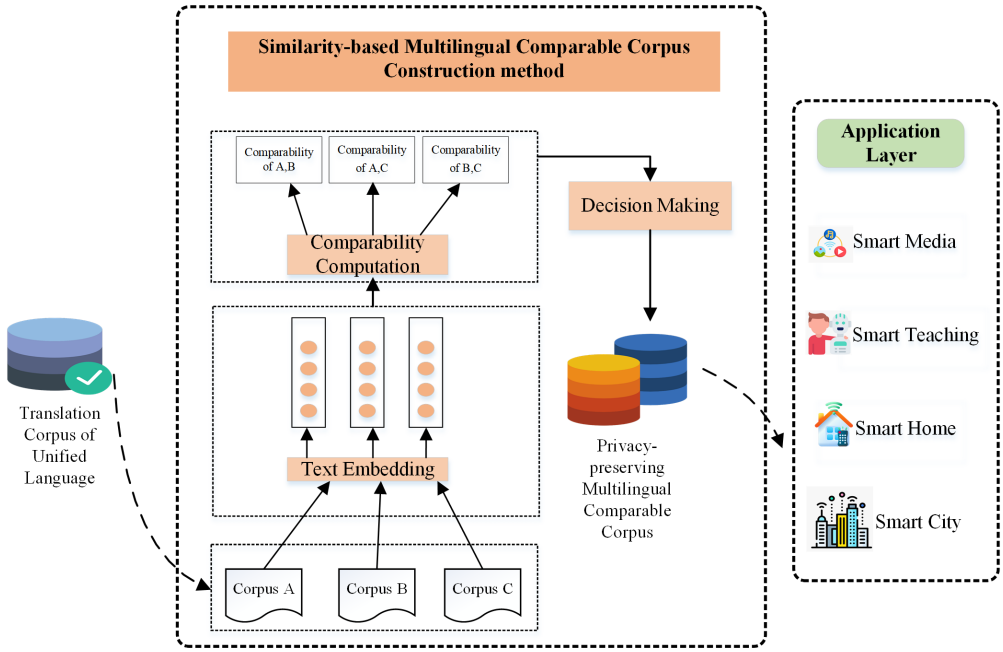

Figure 2. The construction framework of privacy-preserving multilingual comparable corpus.

3.1. Privacy-Preserving Text Embedding under the Unified Language

When the data access is no longer under the control of the data owners, there are potential threats to data security and privacy. Multilingual scenarios become more and more common, and multilingual data sharing is faced with the risk of being cross-lingual. In this paper, we use C as the unified language description, U_m, T_l raw news have been translated into C documents UC_m and TC_l respectively. In general, C_k, UC_m and TC_l are composed of a number of sentences, and each sentence is composed of many words. In order to ensure the security of the source data and calculate the similarity between a triplet (C_k, U_m, T_l) of a raw news data collection, we first embed this triplet (C_k, U_m, T_l) into the text vector space of the unified language description [46]. We cascade the sentences in the text in accordance with their physical location in the document to form a sequence of text in words. And, CW_k, UCW_m and TCW_l can then be described by Equations (1)–(3).

$$C_k \xrightarrow{Identify} C_k \xrightarrow{Concatenate} CW_k = \left\{W_1^{C_k}, W_2^{C_k}, \ldots\ldots W_{MN}^{C_k}\right\} \tag{1}$$

$$U_m \xrightarrow{Translate} UC_m \xrightarrow{Concatenate} UCW_m = \left\{W_1^{UC_m}, W_2^{UC_m}, \ldots\ldots W_{MN}^{UC_m}\right\} \tag{2}$$

$$T_l \xrightarrow{Translate} TC_l \xrightarrow{Concatenate} TCW_l = \left\{W_1^{TC_l}, W_1^{TC_l}, \ldots\ldots W_{MN}^{TC_l}\right\} \tag{3}$$

In the Equation above, where MN stands for the maximum number of words in C_k, UC_m and TC_l, the insufficient part can be padded.

Given that the vector space embedding [47] matrix of the uniform language description of the text is Q (obtained by training, so that CW_k, UCW_m and TCW_l, can be converted into a set of embedding semantic vectors CE_k, UCE_m and TCE_l by Equations (4)–(6), respectively,

$$CE_k = Q^T \times CW_k \tag{4}$$

$$UCE_k = Q^T \times UCW_m \tag{5}$$

$$TCE_l = Q^T \times TCW_l \tag{6}$$

where $CE_k = \left\{E_1^{C_k}, E_2^{C_k}, \ldots\ldots, E_{QN}^{C_k}\right\}$, $UCE_k = \left\{E_1^{UC_m}, E_2^{UC_m}, \ldots\ldots, E_{QN}^{UC_m}\right\}$, $TCE_l = \left\{E_1^{TC_l}, E_2^{TC_l}, \ldots\ldots, E_{QN}^{TC_l}\right\}$ QN stands for the Columns of the matrix Q.

Through Equations (1)–(6), we embed $C = \{C_1, C_2, \ldots, C_K\}$, $U = \{U_1, U_2, \ldots, U_M\}$ and $T = \{T_1, T_2, \ldots, T_L\}$ which denote the sets of Chinese, Uighur and Tibetan raw news texts, respectively, into the text vector space described by the unified language, forming the set of embedded semantic vector groups as follow:

$$C = \left\{ \begin{array}{c} C_1 \\ C_2 \\ \cdots \\ C_K \end{array} \right\} \xrightarrow{identify\ \&\ concatenate\ \&\ embedding} CE = \left\{ \begin{array}{c} CE_1 \\ CE_2 \\ \cdots \\ CE_K \end{array} \right\}$$

$$U = \left\{ \begin{array}{c} U_1 \\ U_2 \\ \cdots \\ U_M \end{array} \right\} \xrightarrow{translate\ \&\ concatenate\ \&\ embedding} UCE = \left\{ \begin{array}{c} UCE_1 \\ UCE_2 \\ \cdots \\ UCE_M \end{array} \right\}$$

$$T = \left\{ \begin{array}{c} T_1 \\ T_2 \\ \cdots \\ T_L \end{array} \right\} \xrightarrow{translate\ \&\ concatenate\ \&\ embedding} TCE = \left\{ \begin{array}{c} TCE_1 \\ TCE_2 \\ \cdots \\ TCE_L \end{array} \right\}$$

3.2. Calculation of Privacy-Preserving Multilingual Text Comparability

Protecting the privacy and security of the corpus ensures that data are used within a legal, reasonable, and transparent framework. This helps prevent data misuse, protect user rights, and promote sustainable data innovation and research. Based on this, the key issue in constructing a privacy-preserving multilingual comparable corpus and ensuring its security is how to calculate the comparability of different corpus texts. In this paper, the similarity degree is used as an evaluation index for the comparability between two or more different texts. Therefore, the solution of comparability in this paper is essentially oriented toward the calculation of the similarity between C, U and T texts in pairs. The methods and steps adopted are as follows:

Let $p_q^{q_i}$ $(q \in (1,2,3))$ stand for CE_i, UCE_i and TCE_l, where, $q = 1$ stands for CE, $q = 2$ stands for UCE, $q = 3$ stands for TCE. We defined $S_{uv}^{u_i v_i}$ as the union of $p_u^{u_i}$ and $p_v^{v_i}$, $u \neq v \in q$. Supposing that the length of the $S_{uv}^{u_i v_i}$ is L_{uv}, $t_y (y \in (1,2,\ldots,L_{uv}))$ represents the number yth elements of $S_{uv}^{u_i v_i}$, the count vector of $p_v^{v_i}$ and $p_v^{v_i}$ can be defined as follows:

$$F_u^{S_{uv}^{u_i v_i}} = \left[np_u^{u_i}(t_1), np_u^{u_i}(t_2) \ldots\ldots np_u^{u_i}(t_{L_{uv}}) \right] \tag{7}$$

$$F_v^{S_{uv}^{u_i v_i}} = \left[np_v^{v_i}(t_1), np_v^{v_i}(t_2) \ldots\ldots np_v^{v_i}(t_{L_{uv}}) \right] \tag{8}$$

In $S_{uv}^{u_i v_i}$, where $np_u^{u_i}(t_y)$ stands for the frequency of the number yth character occurring in $p_u^{u_i}$; $np_v^{v_i}(t_y)$ stands for the frequency of the number yth character occurring in $p_v^{v_i}$. Then, the similarity value $S_{p_u^{u_i}}^{p_v^{v_i}}$ [35,48,49] between CE_i, UCE_i and TCE_i can be calculate by Equation (9):

$$S_{p_u^{u_i}}^{p_v^{v_i}} = \frac{F_u^{S_{uv}^{u_i v_i}} \odot F_v^{S_{uv}^{u_i v_i}}}{\left\| F_u^{S_{uv}^{u_i v_i}} \right\| \times \left\| F_v^{S_{uv}^{u_i v_i}} \right\|} \tag{9}$$

where $\odot$ means vector inner product operating, $\|.\|$ means vector norm operating.

3.3. The Decision Making Mechanism of Privacy-Preserving Multilingual Comparability

Based on the privacy-preserving of corpus data sharing, using comparable pairs can keep the true value of a corpus and ensure the security of the source data. In order to figure out proper comparable pairs for private data sharing, this section aims to point out a decision making mechanism of privacy-preserving multilingual comparability between different news texts. According to Equations (7)–(9), the similarity $S_{p_u}^{p_v}$ between CE_i , UCE_i

and TCE_i can be calculated, and $0 \leq S_{p_u}^{p_v} \leq 1$. When $0 \leq S_{p_u}^{p_v} \leq 1$, it means that the documents p_u and p_v are not similar, in other words, they are not comparable; when $0.5 \leq S_{p_u}^{p_v} \leq 1$, it means that the documents p_u and p_v are similar, in other words they are comparable; when $S_{p_u}^{p_v} = 0.5$, it is difficult to decide whether the documents p_u and p_v are similar or not. Therefore, the similarity degree of 0.5 becomes a key point to determine whether two documents are similar or not. In this study, the impossibility minimization principle is adopted, and the similarity degree of 0.5 is chosen as the threshold value for determining the comparability of two documents; i.e., when the similarity degree is higher than 0.5, the two documents are determined to be comparable; otherwise, they are not comparable. The corresponding decision making process is as follows:

1. Choose one piece of Chinese news i from CE randomly, and search the proper news UCE_j that matches the maximum similarity in the corpus UCE

$$S_{p_u^{u_i}}^{p_v^{p_j}} = \max_{v_b \in (1,2 \ldots M)} S_{p_u}^{p_v^{p_b}} \tag{10}$$

2. Aim at the number jth piece of news in UCE, search for the most similar news TCE_l in TCE.

$$S_{p_u^{u_i}}^{p_v^{v_l}} = \max_{v_b \in (1,2 \ldots M)} S_{p_u^{u_i}}^{p_v^{v_b}} \tag{11}$$

3. Calculate the similarity value of the number ith piece of Chinese news and the number lth Tibetan news,

$$S_{p_u^{u_i}}^{p_v^{v_l}} = \frac{F_u^{S_{uv}^{u_i v_l}} \odot F_v^{S_{uv}^{u_i v_l}}}{\left\| F_u^{S_{uv}^{u_i v_l}} \right\| \times \left\| F_v^{S_{uv}^{u_i v_l}} \right\|} \tag{12}$$

4. The rule of similarity decision making mechanism: $S_{p_u^{u_i}}^{p_v^{v_j}} \cap S_{p_u^{u_i}}^{p_v^{v_l}} \cap S_{p_u^{u_i}}^{p_v^{v_l}} \leq 0.5$, if it fits the conditions, then skip to procedure (6);
5. The number ith piece of news in news corpus C, the number jth news in news corpus U and the number lth piece of news in news corpus T form a comparable corpus pair and are entered into the final C-U-T comparable corpus. Delete the ith piece of news from C, the jth news from Uand the lth news from T.
6. Finish repeating procedure (1)–(5) until all news in the corpus are comparable.

We outline the main procedures of our proposed method in the following Algorithm 1:

Algorithm 1 Algorithm describing the forward steps of constructing multilingual comparable corpus.

Input: Root directory path rootDir
Output: A list of qualified file groups

```
Initialize resultList as an empty list
dirList ← Get all sub-directory paths within rootDir
Assign C, T, U ← dirList[0], dirList[1], dirList[2]
for all file fc in directory C do
    for all directory dir in [T, U] do
        Initialize qualifiedList as an empty list
        for all file f in directory dir do
            similarity ← Compute cosine similarity between contents of fc and f
            if similarity > 0.5 then
                Append f to qualifiedList
```

Algorithm 1 *Cont.*

```
            end if
        end for
        if dir is T then
            q ListCT ← qualifiedList
        else
            q ListCU← qualifiedList
        end if
    end for
    for all file ft in qListCT do
        for all file fu in qListCU do
            s ← Compute cosine similarity between contents of ft and fu
            if s > 0.5 then
                Compute similarities Vc_t, Vc_u, and Vt_u for pairs (fc, ft), (fc, fu), and (ft, fu) respectively
                resultTuple ← (fc, ft, fu, Vc_t, Vc_u, Vt_u)
                Append resultTuple to resultList
            end if
        end for
    end for
end for
Record resultList to a file
```

4. Experiments and Results

4.1. Experimental Environment

The hardware environment for the experiments consists of a server with a 32G Tesla V100 graphics card, two Xeon 4210R CPUs and 128 G of RAM. The software environment used for the experiments and evaluation is as follows: Ubuntu 18.04 for the server operating system, Scrapy for the crawler architecture, Beautifulsoup for the HTML parsing and analysis tool, Niutrans (https://niutrans.com/, accessed on 30 August 2022) for the translation toolkit, Simtext for the similarity calculation toolkit. The development software is Python 3.6.

4.2. Data Preparation for Privacy-Preserving Multilingual Comparable Corpus

IoT provides massive data support to implement smart services related to corpora, like new broadcasting that often aggregates source data from numerous sensors, such as disaster warnings, disease surveillance data, etc. Due to considerations of data security, currently, there is limited availability of open data for individuals to access news data from mobile sensors, satellite remote sensing sensors, and wearable sensor devices [19]. With the intelligent development of mainstream media, there has also been effective coordination between AI, IoT, and other technologies in the collection of news data, enabling real-time intelligent mining and collection. However, unprocessed IoT news texts frequently encounter challenges like shorter text length and sparse features. Therefore, in order to enhance the capability of the proposed model and ensure the security and standardization of the data, the news data in this study is sourced and collected from news data generated by IoT devices. News texts are public, authentic, pertinent, timely and declarative in style to facilitate the processing of information in the text. In addition, compared to literature, news texts are larger and more thematically rich, have a clear timeline and location source, and are more likely to be of common interest to speakers of different languages. Based on these features, this study chooses news texts as its research material. To ensure the authority, authenticity and comprehensiveness of the corpus for this study, Chinese, Uighur and Tibetan online news published by *People's Daily Online*, *Tianshan News* and *China Tibetan News* are selected as the resource of the corpus.

The data in a news corpus may contain sensitive information, such as personal information, location, political or economic data. If this data is leaked or misused, it may lead to

potential privacy protection problems. Therefore, protecting the privacy and security of the corpus during data acquisition, storage, and processing is necessary to prevent these risks. Although the data in our research were collected from news websites, we still need to be concerned about security and privacy protection. The final raw corpus results consist of 24,571 news texts in Chinese; 18,858 news texts in Uighur; and 26,388 news corpora in Tibetan, and a total of 69,817 news texts, including six categories such as economy, society, emergencies, and so on. Currently, there are 3534 pairs of Chinese–Uighur–Tibetan comparable corpora available. Given the limited availability of multilingual comparable corpora, it is impractical to directly evaluate and compare the quality of diverse ethnic multilingual comparable corpora; in this study, we take 50 pairs of privacy-preserving CUTCC comparable corpora randomly as the experimental data set.

4.3. Metrics

In order to evaluate the validity of the construction methods of a privacy-preserving multilingual comparable corpus, we take [50] as comparative experiments, and we employ accuracy, recall, and F1 score as evaluation metrics for the models. As for the quality our multilingual comparable corpus, we take Ning's study [43] as a comparative example to verify.

The main idea of the corpus word feature evaluation method is to evaluate the overall comparability based on words extraction as their features, with reference to accuracy rate, recall rate and *F*-value as evaluation metrics, where accuracy represents the proportion of co-occurring (correct) words among all words; recall is the proportion of co-occuring (correct) words among all words in the corpus; and *F*-value is the summed average of the accuracy and recall rates. The specific corresponding Equation is as follows:

$$P = \frac{\text{correctly words}}{\text{the number of all extract words}} \tag{13}$$

$$R = \frac{\text{correctly words}}{\text{total words}} \tag{14}$$

$$F = \frac{P \times R \times 2}{P + R} \tag{15}$$

From the perspective of privacy-preservation, comparable pair can be regarded as one of the methods of substitution in the process of data sharing de-identification. Comparability is an important metrics for the internal evaluation of a comparable corpus, where "comparability" is quantified as the degree of similarity between specific texts in terms of title, time, genre, grammatical morphology, semantic content and semantic features [17]. Given the dearth of multilingual comparable corpus, particularly for low-resource languages, direct assessment and comparison of different multilingual comparable corpora becomes challenging. The aim of this study is to assess the similarity by calculating between randomly selected pairs of the *CUTCC* using the methodology proposed in [31]. In order to conduct a comparative experiment, a sample of 50 *CUTCC* comparable pairs is utilized.

4.4. The Realization of Privacy-Preserving Multilingual News Comparable Corpus

4.4.1. Chinese–Uighur–Tibetan Data Collecting and Processing

In Chinese context, an ethnic area is always faced with multilingual and multicultural occasions in order to solve cross-lingual communication and security protection of information exchange problems, so in our study, we take Chinese, Uighur and Tibetan as an example to describe our construction method. To ensure the source data's value and safety, the multilingual news source is stable from the official website, including *People's Daily Online*, *Tianshan News* and *China Tibetan News* and so on. The specific collection process and steps are as follows (see also in Figure 3):

Firstly, the data capture crawler software (DCCS) is developed under the Scrapy framework; the pre-defined news websites used this tool to obtain all the web pages that

meet the time interval requirement (the time interval of the news reports selected in this paper is 1 January 2019–1 August 2023.)

Using the BeautifulSoup tool in the Python library to parse and analyze the HTML of the obtained web pages, extracting core information such as title, content, time and location.

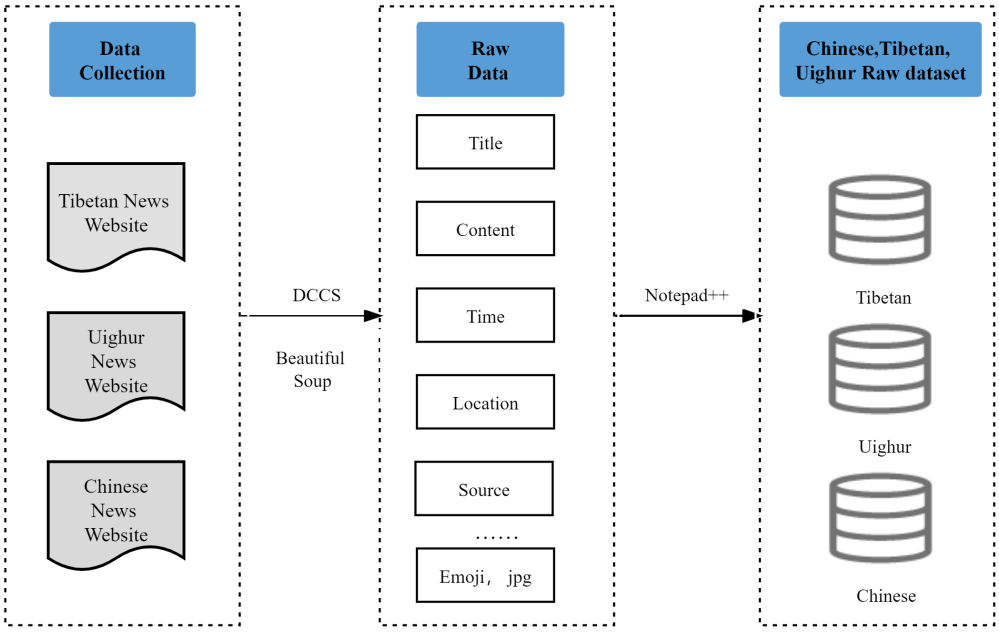

Figure 3. The main steps of news data collection.

Preprocess the acquired text, mainly including corpus standardization, using Notepad++ software to convert the original corpus to UTF-8 encoding format. The corpus cleaning session mainly removes data tags, emojis and other redundant symbols other than text in the corpus. The preprocessing of the corpus is completed to form the original datasets of the Chinese, Uighur, and Tibetan news corpora, respectively denoted as C, U, and T. The titles are used as the document names and chapters as alignment units. The specific data formats and included information are shown in Table 3.

Table 3. Original Chinese–Uighur–Tibetan news data set format.

Storage Format	Title	Source	Content	Time	Location
.txt	-	http://politics.people.com.cn/n1/2021/0119/c1001-32005216.html	-	- Year -Month -Day: -	-

Based on the construction idea proposed in this study, the problem of preserving privacy and calculating the comparability of cross-language texts can only be solved by translating the original Uighur and Tibetan news corpus documents in into corresponding Chinese translated texts, so further processing and analysis of the original corpus data set is required. By means of machine translation and comparable pairs, identification not only avoids the issue of misuse of source language data due to language barriers, but also ensures the usability of the data in the application process. In this study, Niutrans (https://niutrans.com/, accessed on 1 October 2023) is used to translate the original news corpus documents from the Uighur and Tibetan languages into corresponding Chinese translated news corpus texts, denoted as UC and TC, respectively. In order to ensure the high quality of the corpus and minimize information loss resulting from machine translation, this study randomly selected 100 Uighur and Tibetan news texts for the manual evaluation of the machine translation. The results demonstrate commendable usability,

indicating that Niutrans can be employed as an experimental machine translation tool with the format of the translated corpus data set.

4.4.2. The Construction Results of Privacy-Preserving Chinese–Uighur–Tibetan Comparable Corpus

After completing the above-mentioned corpus collection and processing, the following steps are constructed, and the results of this study are derived according to the design methodology and ideas of this study:

Using Equations (1)–(6) to cascade documents C, UC and TC, as well as semantically embed, we form the set of semantic embedding vectors CE, UCE and TCE, which are under a unified language description.

According to Equations (7)–(9), we calculate the similarity values between CE, UCE and TCE in pairs.

Based on the previous similarity results among CE, UCE and TCE, the comparable corpus can be constructed by using Equations (10)–(12) to make decisions on their comparability between corpus.

Through the above process and processing steps, the construction of a privacy-preserving comparable corpus of Chinese–Uighur–Tibetan news is completed. The corpus consists of two aspects: the first is the original corpus data set, which contains three folders, each containing several documents in txt format; the second is the comparable corpus data set, which is mainly stored in the form of comparable corpus pairs, i.e., Chinese, Uighur and Tibetan news texts satisfying the comparable relationship and are formed into a searchable triad according to their serial numbers in the original corpus data set. A searchable triad, a triplet (C_3, U_3, T_3), is formed by the Chinese third news item, the Uighur third news item and the Tibetan third news item, and each comparable pair is numbered in the order of calculation and decision generation. There may be a one-to-one correspondence between them or a one-to-many relationship. The specific data formats are shown in Table 4 below.

Table 4. *CUTCC* data format.

ID	Language	Title	Content	Location	Time
CUTCC3	Chinese	$\left\{ \begin{matrix} C_1, C_2, \\ \ldots, C_K \end{matrix} \right\}$	$\left\{ \begin{matrix} C_1, C_2, \\ \ldots, C_K \end{matrix} \right\}$	-	- Year - Month - Day: -
	Uighur	$\left\{ \begin{matrix} UC_1, UC_2, \\ \ldots, UC_M \end{matrix} \right\}$	$\left\{ \begin{matrix} UC_1, UC_2, \\ \ldots, UC_M \end{matrix} \right\}$	-	- Year - Month - Day: -
	Tibetan	$\left\{ \begin{matrix} TC_1, TC_2, \\ \ldots, TC_L \end{matrix} \right\}$	$\left\{ \begin{matrix} TC_1, TC_2, \\ \ldots, TC_L \end{matrix} \right\}$	-	- Year - Month - Day: -

For the convenience of facilitating the reading and understanding of a wider audience, this study expresses the sample of privacy-preserving multilingual comparable corpus pairs, with the source languages being Tibetan, Uyghur, and Chinese in English. However, the actual texts presented in the corpus still consist of the source texts in multiple languages and their corresponding Chinese translations. The specific examples can be seen in Table 5.

In order to demonstrate the feasibility of the proposed research method, a limited time period has been devoted to constructing a multilingual comparable corpus. Currently, there are 3534 pairs of Chinese–Uighur–Tibetan multilingual comparable corpora available.

4.5. Evaluation Results

To protect the privacy and security of the corpus, commonly used methods include anonymization and de-identification, in addition to which the use of data anonymization methods, such as deleting or replacing personal identifying information, can reduce the risk of data identification. In this study, we aim to evaluate the quality of the corpus that is realized by using the privacy-preserving comparable corpus construction method. External evaluation refers to the indirect measurement of corpus quality by measuring

Table 5. CUTCC comparable pair (sample).

ID	Language	Title	Content	Location	Time
	Chinese	A 5.1 magnitude earthquake occurred in Taitung County, Taiwan, with a focal depth of 10 km.	According to the China Earthquake Networks Center, it has been officially determined that a 5.1-magnitude earthquake occurred in Taitung County, Taiwan at 09:56 on 4 April, with a focal depth of 10 km.	Taitung County, Taiwan, China	4 April, 09:56.
CUTCC3	Uighur	A 5.1 magnitude earthquake occurred in Taiwan County, with a focal depth of 10 km)	According to the China Seismological Network, a 5.1-magnitude earthquake was officially determined to have occurred in Taitung County, Taiwan at 09:56 on 4 April. The earthquake had a focal depth of 10 km.	Taitung County, Taiwan Province, China	9:56 on 4 April,
	Tibetan	The focal depth of the M 5.1 earthquake in Taitung County, Taiwan Province is 10 km	According to China Seismological Network, China Seismological Network officially determined that at 09:56 on 4 April, a 5.1-magnitude earthquake occurred in Taitung County, Taiwan Province, with a focal depth of 10 km.	Taitung County, Taiwan, China	At 9:56, 4 April.

the usefulness of the corpus in specific applications, while internal evaluation refers to the direct measurement of the corpus using its internal features, including lexical features, word sequence features, and linguistic morphological features [42,44]. In this study, we compared our proposed model with the baseline model, and the evaluation results are as follows:

Due to the lack of privacy-preserving multilingual comparable corpus datasets, it is hard to directly assess and compare the quality of multilingual comparable corpora among different ethnic groups [50], so we make use of the analogy method with other existing bilingual corpus. We compare the *CUTCC* model with the benchmark model to measure the usability of the *CUTCC* built in this research. As shown in Table 6, the evaluation results showed that the accuracy rate of the *CUTCC* reached 77%, the recall rate was 34% and the value reached 47.17%, which also reached a good level of quality compared to the other bilingual corpus. The internal analysis of linguistic features not only confirms the existence of comparable relationships between the Chinese, Uighur and Tibetan news corpus, but also indicates that the comparability of this corpus is high and stable, leading to better use and a wider application range. Analysis of the comparative experiments results in Table 6 also reveals that the R value of the proposed model performs lower than baseline model, which may owe to the original texts with varying lengths. The Chinese–Uighur–Tibetan news collected for this study mainly from IoT news, which are typically concise [23] and characterized by sparse content. Additionally, there are variations in the distribution of topics, with IoT news primarily covering themes related to natural disasters and transportation. Moreover, the multilingual news corpus also exhibits uneven distribution. In future improvement efforts, this study will comprehensively consider the collected raw data and enhance the capabilities of the model.

Table 6. The evaluation results.

Metrics	P	R	F
Baseline	43.18%	38%	40.42%
CUTCC	77%	34%	47.17%

As for the comparability parameter, due to the small number of open datasets that can be directly used to evaluate multilingual comparable corpus at this stage, this study uses the internal evaluation method to evaluate the Chinese–Uighur–Tibetan News Comparable Corpus, which is conducted through the word feature description of the corpus. The quality

of the corpus was assessed on the basis of the Jaccard similarity value, including seven static metrics, and the results are shown in Table 7 below:

Table 7. The Jaccard metrics evaluation results.

Metrics		Jaccard
Dispersion	Variance	0.0123
	Standard Deviation	0.1109
	Range	0.376
Central Tendency	Mean	0.704
	Median	0.734
	Mode	0.684
Location	Max	0.922
	Min	0.546

In terms of privacy-preserving, during the data sharing process, how to ensure the quality of data as well as its security is an essential part of the evaluation. The Jaccard similarity value is commonly used to measure the similarity of two collections, with a higher Jaccard value indicating a higher level of similarity. As can be seen in Table 7 above, by calculating the intra-corpus Jaccard index, the mean value of 0.704, median value of 0.734 and mode value of 0.684 indicate that the overall corpus has reached a high level of comparability. From the perspective of statistics of data dispersion, the variance, standard deviation and range value of evaluation comparable corpus pairs' Jaccard similarity (0.0123, 0.1109, 0.376) is small, indicating that the level of comparability between comparable pairs within the overall corpus is not only consistently high, but also tends to be stable, thus making the usability of this comparable corpus better. From the perspective of the distribution of the evaluation data's location, the highest similarity value reaches 0.922 for the Chinese–Uighur–Tibetan comparable corpus pairs, indicating that there are texts with high correspondence in the comparable corpus pairs, which can be used as potential multilingual parallel corpora, as well as potential multilingual inter-translation texts [42]. Through further processing and research on the low-resource corpus, they can be expanded to a certain extent to achieve data widening.

The above assessments show that, despite the difficulty of constructing a privacy-preserving multilingual comparable corpus, the corpus obtained in this study has achieved a high proportion of comparability among texts, which to a certain extent supports the efficiency of our proposed method in multilingual comparable corpus constructing and comparable relationship decision mechanisms proposed in this study. Also, the security and privacy of the corpus have been protected. On the basis of privacy protection, in order to better compensate for the lack of ethnolinguistic databases, we will further develop a web page system in the future not only for remote reviewers, but also for related scholars to conduct further research. Furthermore, future work will be undertaken to develop a monitoring corpus and a refresh multilingual corpus in real time for the further use of more IoT devices and applications.

5. Conclusions

This paper proposes a method for constructing a provable privacy-preserving multilingual news comparable corpus based on web crawler technology. We apply a unified vector space model to calculate comparability and select high-quality comparable pairs. The proposed method is feasible in practice and has resulted in the construction of a privacy-preserving Chinese–Uighur–Tibetan news comparable corpus with a high level of comparability as a sample. To reinforce information security during the corpus construction and data sharing, the method incorporates a translation process and comparable pair substitution, which provides a way of protection and ensures the privacy of multilingual raw information, thus addressing the identified threats. The construction of a comparable corpus of Chinese–Uighur–Tibetan news is not only important for the study of

low-resource languages like ethnic language translation and natural language processing, but will also be of great significance for the privacy-preserving multilingual language service and IoT applications.

However, this study still has some disadvantages. In the construction, the machine translation part requires manual proofreading and expert intervention to some extent, which can further enhance its intelligence and accuracy. In addition, the calculation of comparability between trilingual or even multilingual data involves a large amount of computation, resulting in a lengthy construction time. These disadvantages will also become the focus of our future work. In future work, the issues of data privacy and security will receive widespread and significant attention. With the help of privacy-preserving multilingual corpus resources, IoT applications and devices will be able to provide much more multilingual user facility, assistance and comfort in complex situations and keep the security of information by enhancing the quality and variety of generated data.

Author Contributions: Conceptualization, Y.W. and S.D.; methodology, C. and Y.W.; software, S.D.; validation, S.D., Y.W. and C.; formal analysis, Y.W. and C.; resources, S.D.; writing—original draft preparation, S.D., Y.W. and C.; writing—review and editing, S.D. and Y.W.; funding acquisition, Y.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, grant number 2020YFB1406702-3.

Data Availability Statement: The data that support the findings of this study are available on request from the corresponding author, 20400161@muc.edu.cn, upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Rock, L.Y.; Tajudeen, F.P.; Chung, Y.W. Usage and impact of the internet-of-things-based smart home technology: A quality-of-life perspective. *Univers. Access Inf. Soc.* **2022**, 1–20. [CrossRef] [PubMed]
2. Bin, G.; Sicong, L.; Yan, L.; Zhigang, L.; Zhiwen, Y.; Xingshe, Z. AIoT: The Concept, Architecture, and Key Techniques. *Chin. J. Comput.* **2023**, *46*. Available online: https://kns.cnki.net/kcms2/article/abstract?v=rCMvAF-4El1WLvIjsXZvAiChQ0k3XL_bsnLH7YPUPymadeQl07Yn4l2QCxVCT00_44fCKwOqV3BqfGYLToQHOBA5_7c8GU109AwCbRghrzgOcLqM8RjBiYu-a3zDXmea9Atwq5h28dVtTYsbmZu0sQ==&uniplatform=NZKPT&language=CHS (accessed on 1 October 2023).
3. O'Shaughnessy, P.; Lin, Y.X. Privacy Protection Practice for Data Mining with Multiple Data Sources: An Example with Data Clustering. *Mathematics* **2022**, *10*, 4744. [CrossRef]
4. Aljumah, A.; Ahanger, T. Blockchain-Based Information Sharing Security for the Internet of Things. *Mathematics* **2023**, *11*, 2157. [CrossRef]
5. Liang, K.; Zhou, B.; Zhang, Y.; He, Y.; Guo, X.; Zhang, B. A Multi-Entity Knowledge Joint Extraction Method of Communication Equipment Faults for Industrial IoT. *Electronics* **2022**, *11*, 979. [CrossRef]
6. Pilán, I.; Lison, P.; Øvrelid, L.; Papadopoulou, A.; Sánchez, D.; Batet, M. The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. *Comput. Linguist.* **2022**, *48*, 1053–1101. [CrossRef]
7. He, M.; Li, Y. Application of Big Data Technology in News Media Scene Visualization Based on Internet of Things (IoTs). *Math. Probl. Eng.* **2022**, *2022*, 5508125. [CrossRef]
8. Gaimei, G.; Xu, S.; Chunxia, L.; Weichao, D.; Na, W. A Blockchain-based Method for Privacy Protection of Medical Data. *J. Comput. Appl. Res.* **2023**, 1–7. [CrossRef]
9. Zhong, Z.; Zhang, G.; Yin, L.; Chen, Y. Description and Analysis of Data Security Based on Differential Privacy in Enterprise Power Systems. *Mathematics* **2023**, *11*, 4829. [CrossRef]
10. Baker, M. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target* **1995**, *7*, 223–243. [CrossRef]
11. Xu, H.; Jiang, M.; Lin, J.; Huang, C.R. Light verb variations and varieties of Mandarin Chinese: Comparable corpus driven approaches to grammatical variations. *Corpus Linguist. Linguist. Theory* **2020**, *18*, 145–173. [CrossRef]
12. Wang, B. Feature Extraction Method of Machine Translation Equivalent Pairs in Chinese-English Comparable Corpus based OCR Recognition. In Proceedings of the 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 3–5 June 2021; pp. 899–902. [CrossRef]
13. Dominic, P.; Purushothaman, N.; Kumar, A.S.A.; Prabagaran, A.; Blessy, J.A.; John, A. Multilingual Sentiment Analysis using Deep-Learning Architectures. In Proceedings of the 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirenelveli, India, 23–25 January 2023; pp. 1077–1083. [CrossRef]

14. Katsumata, S.; Komachi, M. Towards Unsupervised Grammatical Error Correction using Statistical Machine Translation with Synthetic Comparable Corpus. *arXiv* **2019**, arXiv:1907.09724.
15. Goyal, V.; Kumar, A.; Lehal, M. Document Alignment for Generation of English-Punjabi Comparable Corpora from Wikipedia. *Int. J. E-Adopt.* **2020**, *12*, 42–51. [CrossRef]
16. Huajun, L.; Kaiyue, W. The media industry's format innovation, relationship reconstruction, and development path in the era of intelligent IoT. *J. Lover* **2022**, *4*, 10–14. [CrossRef]
17. Li, J.; Xie, L.; Chen, Z.; Shi, L.; Chen, R.; Ren, Y.; Wang, L.; Lu, X. An AIoT-Based Assistance System for Visually Impaired People. *Electronics* **2023**, *12*, 3760. [CrossRef]
18. Tang, X.; Zhu, L.; Shen, M.; Peng, J.; Kang, J.; Niyato, D.; Abd El-Latif, A. Secure and Trusted Collaborative Learning Based on Blockchain for Artificial Intelligence of Things. *IEEE Wirel. Commun.* **2022**, *29*, 14–22. [CrossRef]
19. Yujie, L. Reflection on the Communication Mechanism and Media of Wearable Smart Devices in the News Field. *Publ. Angle* **2020**, *15*, 63–65. [CrossRef]
20. Tang, X.; Liao, D.; Shen, M.; Zhu, L.; Huang, S.; Li, G.; Man, H.; Xu, J. Confidence-aware Sentiment Quantification via Sentiment Perturbation Modeling. *IEEE Trans. Affect. Comput.* **2023**, 1–15. [CrossRef]
21. Tang, X.; Shen, M.; Li, Q.; Zhu, L.; Xue, T.; Qu, Q. PILE: Robust Privacy-Preserving Federated Learning Via Verifiable Perturbations. *IEEE Trans. Dependable Secur. Comput.* **2023**, *20*, 5005–5023. [CrossRef]
22. Shuman, W.; Aiping, L.; Liguo, D.; Jia, F.; Yongle, C. BTM-based IoT service discovery method. *J. Comput. Appl.* **2020**, *40*, 459–464. [CrossRef]
23. Yimei, W. Content Production Strategy and Practice of Satellite News. *Youth J.* **2022**, *2*, 70–72. [CrossRef]
24. Ruslan, A.; Jusoh, A.; Asnawi, A.L.; Othman, M.; Abdul Razak, N.I. Development of multilanguage voice control for smart home with IoT. *J. Phys. Conf. Ser.* **2021**, *1921*, 012069. [CrossRef]
25. Sayakkara, A.; Le-Khac, N.A. Electromagnetic Side-Channel Analysis for IoT Forensics: Challenges, Framework, and Datasets. *IEEE Access* **2021**, *9*, 113585–113598. [CrossRef]
26. Iliev, Y.; Ilieva, G. A Framework for Smart Home System with Voice Control Using NLP Methods. *Electronics* **2022**, *12*, 116. [CrossRef]
27. Zhang, Q.; Xiang, Z. Improvement of culture media efficiency in Internet of Things based on global numerical ant colony algorithm. *Pers. Ubiquitous Comput.* **2020**, *24*, 347–361. [CrossRef]
28. Wei, P. Research on the Construction Technology of Tibetan-Chinese Bilingual Comparable Corpus Based on Web. Master's Thesis, Minzu University of China, Beijing, China, 2015.
29. Langlois, D.; Saad, M.; Smaïli, K. Alignment of comparable documents: Comparison of similarity measures on French–English–Arabic data. *Nat. Lang. Eng.* **2018**, *24*, 677–694. [CrossRef]
30. Wei, Y. Construction, evaluation and application prospects of Russian-Chinese news comparable corpus. *J. PLA Univ. Foreign Lang.* **2017**, *40*, 8.
31. Lianfu, Z.; Zuowens, T. A Privacy Preservation Method for Multi-Modal Medical Data in Federated Learning. *Comput. Sci.* **2023**, *50*, 933–940.
32. Qihui, T.; Lanjiang, Z.; Chang, L. Textual feature based bilingual sentence similarity measure between Chinese and Lao. *J. Chin. Inf. Process.* **2022**, *35*, 64–72. [CrossRef]
33. Hongjun, W.; Shuicai, S.; Shiwen, Y.; Shibin, X. Cross-language similar document retrieval. *J. Chin. Inf. Process.* **2007**, *21*, 8.
34. Xing, T.; Jin, Z.; Zuping, Z. Jaccard text similarity algorithm based on word embedding. *Comput. Sci.* **2018**, *45*, 186–189.
35. Xiaoli, D.; Shifeng, L.; Daqing, G. NLP-based text similarity detection method. *J. Commun.* **2021**, *42*, 173–181. [CrossRef]
36. Xunyu, L.; Cunli, M.; Zhengtao, Y.; Shengxiang, G.; Zhenhan, W.; Yafei, Z. Chinese-Burmese comparable document acquisition based on topic model and bilingual word embedding. *J. Chin. Inf. Process.* **2021**, *35*, 88–95.
37. Weizhen, Z.; Shuang, R. A Study on the Technology System of Railway Data Security and Privacy Protection. *Railw. Comput. Appl.* **2023**, *32*, 45–50. [CrossRef]
38. Lufang, L.; Bo, L.; Peng, C.; Linghan, Z.; Bing, W. Bilingual lexicon extraction based on word vector and comparable corpus. *Comput. Sci. Eng.* **2018**, *40*, 368–373.
39. Panlu, C. 6G, Semantic Communication, and Future Models of Journalism and Communication: A Digital Journalism Perspective. *J. Guangzhou Univ. (Soc. Sci. Ed.)* **2022**, *21*, 5–16.
40. Wang, X. The Impact of IoT on News Media in the Smart Age. *Mob. Inf. Syst.* **2022**, *2022*, 2238233. [CrossRef]
41. Zhang, J.; Tao, D. Empowering Things With Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things. *IEEE Internet Things J.* **2020**, *8*, 7789–7817. [CrossRef]
42. Nwanakwaugwu, A.; Matthew, U.; Okey, O.; Kazaure, J.; Ubochi, C. News Reporting in Drone Internet of Things Digital Journalism: Drones Technology for Intelligence Gathering in Journalism. *Int. J. Interact. Commun. Syst. Technol.* **2023**, *12*, 22–42. [CrossRef]
43. Ning, S.; Yan, X.; Nuo, Y.; Zhou, F.; Xie, Q.; Zhang, J. Chinese-Khmer Parallel fragments Extraction from Comparable Corpus Based on Dirichlet Process. *Procedia Comput. Sci.* **2020**, *166*, 213–221. [CrossRef]

44. Dalian Minzu University. *A Method of Constructing a Parallel Corpus of Chinese-English-Mongolian-Tibetan Victorian*; No. 18 Liaohe West Road; Dalian Economic and Technological Development Zone: Dalian, China, 2022. Available online: https://kns.cnki.net/kcms2/article/abstract?v=rCMvAF-4El0GzZ5X9eGvD8ATcYVVIhH19Df_FMaey6NT0D6YpiI9mvcbcPRDuaZLEq2D8RuHPzmRu4ofEIF5zqrrtiEJPcM92H-_03dOHzoS-F5_zPhG38gBLu3TwUMlg5y3ac7bkEU=&uniplatform=NZKPT&language=CHS (accessed on 1 October 2023).
45. Lei, C.; Weibin, Y.; Qinyao, S.; Zhi, W.; Chongzhong, Y.; Daowei, L. Research and construction of endangered language spoken corpus-case study on Lizu. *Comput. Eng. Appl.* **2018**, *54*, 234–238. [CrossRef]
46. Cohen, L.; Christopher, M.; Quoc, N. *NBER Working Paper Series*; National Bureau of Economic Research: Cambridge, MA, USA, 2018. Available online: http://www.nber.org/papers/w25084 (accessed on 1 October 2023).
47. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
48. Sa, L. Building and Evaluating Special Domain Comparable Corpus. Master's Thesis, Nanjing University of Science and Technology, Nanjing, China, 2012. Available online: https://kns.cnki.net/kcms2/article/abstract?v=rCMvAF-4El2rI7_R9d-DLHpt8ZdySbER3tlBKhiyUSqwln4Gn3z1b03sy_uDfXRvWb9w07GNk99u14O89yOLdxBTjPClkraUYNU9ae9Lp2TAnRB29l8iY3IPcacXVJZ3JpFEq10E0IgvqTfQ0d-9sQ==&uniplatform=NZKPT&language=CHS (accessed on 1 October 2023).
49. Chengcheng, H.; Lei, L.; Tingting, L.; Ming, G. Approaches of semantic textual similarity. *J. East China Norm. Univ. (Nat. Sci. Ed.)* **2020**, *5*, 95–112.
50. Fei, P.; Ibrahim, T.; Er, A.S.W.; Litipu, M. Construction of Chinese-Uighur comparable corpus for alignment of bilingual technical terms. *J. Xinjiang Univ. (Nat. Sci. Ed.)* **2017**, *34*, 316–321.