Postprint: Research on Multi-Feature-Based Cross-Language Plagiarism Detection Technology

Authors: Liu Gang, Hu Yulin, Li Guangxi

Date: 2018-10-11T00:00:00+00:00

Abstract

To address the problem of detecting bilingual plagiarism, a cross-language plagiarism detection model is proposed. The model comprises cross-language plagiarism classification based on multi-feature selection and cross-language plagiarism detection based on multi-feature correspondence. The method primarily mines common translation features based on the Europeanization phenomenon exhibited by translators during translation, performs further feature selection and feature weight computation, trains a classifier to categorize the presence of cross-language plagiarism behavior, and finally conducts plagiarism confirmation via WordNet. Through experimental comparison and analysis, both classification and detection results are validated, thereby demonstrating the effectiveness and scientific validity of the proposed model.

Full Text

Preamble

Abstract: To address the problem of bilingual plagiarism detection, this paper proposes a cross-language plagiarism detection model. The model comprises cross-language plagiarism classification based on multi-feature selection and cross-language plagiarism detection based on multi-feature correspondence. The method primarily mines common translation features based on the Europeanization phenomenon that occurs when translators perform translations. After further feature selection and feature weight calculation, a classifier is trained to categorize whether cross-language plagiarism exists. Finally, WordNet is used for plagiarism confirmation. Through experimental comparison and analysis, both classification and detection results are validated, demonstrating the effectiveness and scientific validity of the proposed model.

Keywords: cross-language plagiarism detection; bilingual features; Europeanized grammar

1.1 Plagiarism Classification

Plagiarism is divided into literal plagiarism and intelligent plagiarism. Literal plagiarism is more common and does not deliberately conceal plagiarized content, merely achieving its goal through copy-pasting. Literal plagiarism is further subdivided into three types: (a) exact copying, which involves copying a paragraph or entire article without any modification; (b) similar copying, which involves operations such as insertion, deletion, substitution, sentence splitting or merging before copying; and (c) modified copying, which involves phrase reordering or morphological and syntactic changes, or plagiarism through conceptual induction, summarization, and interpretation. Overall, literal plagiarism involves minimal changes without citing the original work.

Intelligent plagiarism refers to attempts to hide and alter the original article through various means, primarily categorized into three methods: (a) text processing, which changes vocabulary and morphological grammar; (b) translation, which translates from one language to another through automatic translation (exact translation, parallel corpora, etc.) or manual translation without citation, also constituting plagiarism; and (c) idea plagiarism, the most serious form, which involves stealing others' ideas without citation.

1.2 Cross-language Text Similarity Algorithms

Machine translation-based methods represent the most direct and simplest approach for cross-language similarity calculation. They achieve cross-language similarity computation by unifying both languages into the same form for comparison.

Multilingual dictionary-based algorithms primarily perform matching through bilingual dictionary correspondence. These have been applied in both CLIR (Cross-Language Information Retrieval) and CLSD (Cross-Language Plagiarism Detection), initially emerging from CLIR and subsequently developing in CLSD with good results. A typical algorithm is CL-CNG (Cross-Language Character N-Gram) [12]. It should be noted that the CL-CNG algorithm is only suitable for similar languages and not applicable to significantly different languages like Chinese and English.

The most typical algorithm in this category is the Cross-Language Explicit Semantic Analysis (CL-ESA) algorithm, an extension of the ESA algorithm proposed by Martin Potthast et al. in 2008. Before introducing CL-ESA, the ESA algorithm for monolingual semantic similarity analysis must be described. ESA uses Wikipedia as a concept space, represents text vectors using the vector space model, calculates weights using TF-IDF, represents texts through concept

weight lists in the concept space, and computes similarity between two vectors through cosine similarity.

Similarly, CL-ESA [21] extends the ESA algorithm to cross-language scenarios, establishing a concept space based on bilingual Wikipedia with concept alignment between the two languages. The process is shown in Figure 1 [Figure 1: see original paper].

---

2 Cross-language Plagiarism Classification Based on Multi-feature Selection

For cross-language plagiarism, the first step should be determining whether a given article contains cross-language plagiarism. Articles with cross-language plagiarism must be identified before determining which paragraphs or sections contain the plagiarism. To address this issue, this chapter focuses on discovering and selecting effective translation features from Chinese articles with cross-language plagiarism, assigning different feature weights, and constructing a classification model that can categorize given Chinese articles to detect which ones may contain plagiarism and which do not.

2.1 Europeanization Phenomenon and Translationese Issues in English-Chinese Translation

Translationese is the manifestation of Europeanization, referring to translated texts that exhibit Europeanization phenomena or do not conform to Chinese habitual expression patterns, also known as translation tone or translation syndrome. Literature [20] translates this as "translationese." Europeanization, also called Westernization, refers to Chinese language that is overly influenced by European languages, particularly English, in terms of grammar, writing style, or word usage [48]. Europeanized Chinese appears slightly rigid in language expression and word usage and is relatively easy to identify.

Dr. Li Yingyu from Shanghai International Studies University summarized common Europeanized translation manifestations into seven forms: (a) foreign words and affixation; (b) letter word usage; (c) increased conjunctions; (d) flexible word class usage; (e) abuse of auxiliary words, quantifiers, and pronouns; (f) long and redundant sentences; and (g) increased passive voice usage, explicit markers, and singularization tendencies.

Thus, among the many factors influencing mutual English-Chinese language impact, lexical and grammatical influences are relatively significant and constitute the most prominent manifestations distinguishing Europeanized translation. The renowned Chinese linguist Mr. Wang Li dedicated an entire chapter, "Europeanized Grammar," in literature [7] to discussing Europeanization phenomena and criticized some "malicious Europeanization" instances. "Malicious Europeanization" exists not only among those who do not translate professionally

but also among excellent translators who may make oversights, especially when dealing with articles from different fields. Therefore, abstracting translation features to determine whether an article has cross-language plagiarism problems is feasible, and constructing and selecting appropriate translation features is key to building a classification model.

2.2 Feature Selection–Improvement of Chi-square Test

This paper utilizes the chi-square test for preliminary translation feature selection and improves upon CHI based on its shortcomings, aiming to remove features with low frequency and unstable distribution across categories to accurately identify effective features for precise classification.

Let category c_j contain n_j articles, and feature term t_i appear with frequency f_{ik} in each article. The average frequency of feature term t_i in c_j is calculated as shown in Equation (1):

$$\alpha_{ij} = \frac{1}{n_j} \sum_{k=1}^{n_j} f_{ik}$$

Using all articles in the denominator rather than only those containing feature term t_i prevents the scenario where low-frequency words appear frequently in a small subset of articles but not in the vast majority, which would increase the frequency value and reduce discrimination for rare words. Consequently, the frequency difference for feature term t_i is defined and normalized as:

$$\alpha_i = \max_{j_1,j_2}(\alpha_{ij_1}, \alpha_{ij_2}) - \min_{j_1,j_2}(\alpha_{ij_1}, \alpha_{ij_2})$$

This constrains the value to the interval [0,1]. A larger frequency difference indicates stronger discriminative capability. Equation (2) addresses the first shortcoming of CHI by introducing α_i to distinguish feature frequency issues.

For the second shortcoming, this paper introduces information entropy. Information entropy measures the uncertainty of random variables, originating from physics as a parameter characterizing the uniformity of energy distribution in space. Let X be a discrete random variable taking finite values with probability distribution:

$$P\{X = x_i\} = p_i, \quad i = 1, 2, ..., n$$

The entropy of random variable X is defined as:

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i$$

with $0 \leq H(X) \leq \log n$. The smaller $H(X)$, the more non-uniform the distribution.

In this paper, the uniform distribution status of feature t_i in specified category c_j must be determined. Let the k-th article in c_j be d_k, then the information entropy of feature term t_i in category c_j is expressed as in Equation (5):

$$H(t_i, c_j) = -\sum_{k=1}^{n_j} \frac{f_{ik}}{c_{ij}} \log \frac{f_{ik}}{c_{ij}}$$

where f_{ik} represents the occurrence count of feature term t_i in article d_k, and c_{ij} is the total occurrence count of feature term t_i in category c_j.

A larger $H(t_i, c_j)$ indicates a more uniform distribution and better feature effectiveness. It is stipulated that if a feature does not exist in a category, $H(t_i, c_j) = 1$. When feature t_i is stable in class c_1 and unstable in class c_2, the value of $H(t_i, c_1) - H(t_i, c_2)$ is larger and better represents the plagiarized class. Thus, for all feature terms t_i, normalization yields:

$$H_i^{hot} = \frac{H(t_i, c_1) - H(t_i, c_2)}{\max(H(t_1, c_1) - H(t_1, c_2), ..., H(t_n, c_1) - H(t_n, c_2))}$$

In summary, a new CHI method is defined as:

$$CHI_{new}(t_i, c) = k_1 P + k_2 \alpha_i + k_3 H_i^{hot}$$

where P is the probability obtained by querying the chi-square distribution critical value table with the $\chi^2(t_i, c)$ value; α_i is the average frequency difference of feature term t_i; and H_i^{hot} is the information entropy of feature term t_i in the category. Both latter terms are normalized, so all three components fall within [0,1]. The parameters k_1, k_2, and k_3 are weights for each factor. A larger $CHI_{new}(t_i, c)$ value indicates higher feature discriminability and effectiveness, while a smaller value indicates lower discriminability.

2.3 SVM Model Training

Based on the Europeanization phenomenon and translationese issues in English-Chinese translation, translation features present in Chinese articles are constructed. According to these translation features, Chinese articles that may contain plagiarism are identified. From another perspective, if Chinese translation features correspond to features in English, the specific plagiarism results can be further determined based on the positions where English-Chinese translation features appear.

This paper employs a non-linear Support Vector Machine as the model, selecting RBF (radial basis function) as the kernel function. The classification decision function to be obtained through learning is:

$$f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*\right)$$

where the RBF kernel is:

$$K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

Each feature in every sentence is weighted to obtain two n-dimensional ordered vectors representing the feature manifestation of n features in the compared Chinese and English paragraphs. The Euclidean distance between these two vectors is calculated as the distance between paragraphs. A shorter distance indicates greater similarity between the two paragraphs.

Example 1: Figures 2 [Figure 2: see original paper] and 3 [Figure 3: see original paper] show a Chinese paragraph and its corresponding English paragraph, respectively.

The specific SVM classification model construction and solution method are described in Algorithm 1.

Algorithm 1: SVM Model Construction and Solution Algorithm Based on Translation Features

Input: Training dataset D and feature set T.

Output: Plagiarism classification model.

1. Select parameter C, replace inner product with RBF kernel to obtain the SVM dual problem.
2. Select optimization variables α_{k_1} and α_{k_2}.
3. Convert the dual problem to the form of Equation (2-21).
4. Initialize $\alpha^{(0)} = 0$, set $k = 0$.
5. **while** there exist variables not satisfying KKT conditions:
   - Update variables $\alpha_{k_1}^{(k+1)}$ and $\alpha_{k_2}^{(k+1)}$ that do not satisfy constraints using the SMO algorithm mentioned in Chapter 2.
   - **if** KKT conditions are satisfied within precision ε:
     - **break**
   - $k = k + 1$
6. **end while**
7. Obtain the optimal solution α^* and update b^*.
8. Return the classification model $f(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*\right)$.

3 Cross-language Plagiarism Detection Based on Multi-feature Correspondence

After constructing the classification model and given a Chinese article, it can be determined whether cross-language plagiarism has occurred. When confirming that an article contains cross-language plagiarism, it is necessary to further identify which specific article has been plagiarized and which paragraphs are involved. This chapter proposes a cross-language plagiarism detection method based on multi-feature correspondence, continuing from the previous chapter. Using the plagiarism candidate set obtained previously, this method further analyzes feature correspondence to confirm specific plagiarized content.

3.2 Plagiarism Result Secondary Filtering Based on Structural Feature Correspondence

This paper selects five structural features for further screening and filtering of the plagiarism candidate set: sentence length, noun length in sentences, verb length in sentences, adjective length in sentences, and adverb length in sentences, as described in Algorithm 2.

Algorithm 2: Paragraph Filtering Algorithm Based on Structural Features

Input: Chinese plagiarism paragraph P, initial plagiarism result set E.

Output: Filtered plagiarism results.

1. **for** each retained paragraph E_j in E:
   - **if** $\|l_n^P - l_n^{E_j}\| > \varepsilon_n$ **or** $\|l_v^P - l_v^{E_j}\| > \varepsilon_v$ **or** $\|l_{adj}^P - l_{adj}^{E_j}\| > \varepsilon_{adj}$ **or** $\|l_{adv}^P - l_{adv}^{E_j}\| > \varepsilon_{adv}$:
     - Mark this paragraph as non-compliant.
   - **else**:
     - Save this plagiarism paragraph.
2. **end for**
3. Put all screened plagiarism results into a map with P as the key.
4. Return map values.

Algorithm 2 provides thresholds for five features. Given a Chinese paragraph and multiple English paragraphs from the initial filtering, paragraphs not meeting the conditions are filtered out. For 1,000 compared paragraph pairs, 749 retained only one suspicious paragraph after two rounds of filtering, with 736 accurately matching the plagiarized paragraph and only 13 mismatches. In the remaining 251 paragraph candidate sets, 24 had no matching suspicious paragraphs after two rounds of filtering, while 227 had multiple matching suspicious paragraphs. Figure 7 [Figure 7: see original paper] shows these results, demonstrating that accuracy has reached 74% at this stage. For the 227 paragraphs matching multiple results, WordNet is needed to identify the specific plagiarized paragraph.

3.3 Final Plagiarism Determination Based on WordNet

After two rounds of filtering, final plagiarism results are obtained. The result may contain only one paragraph, meaning the English paragraph plagiarized by the Chinese paragraph has been found but requires semantic confirmation; or it may contain multiple paragraphs requiring precise identification. Therefore, this section introduces a cross-language text similarity calculation method based on WordNet [22] for final result confirmation.

After noun disambiguation, each noun yields a useful fingerprint sequence, but not all nouns are useful. Some nouns appear with very low frequency and lack typicality, requiring filtering to retain high-resolution fingerprints for similarity calculation.

This paper adopts a method similar to TF-IDF weight calculation for fingerprint selection. Nouns appearing multiple times (with high TF) are retained. For inverse document frequency (IDF) selection, this paper uses the depth of nodes in the WordNet synonym dataset' s tree structure as a filtering condition [11]. Shallower depth indicates weaker semantic meaning. Therefore, fingerprints with depth below 100 are filtered out (all zeros), and the remaining fingerprints are selected for similarity calculation.

After noun semantic hashing, noun disambiguation, and fingerprint selection, formal hash feature sequences are obtained. Let the feature sequences of input text d in language L and input text d' in language L' be $F(d) = \{s_1(\varphi), s_2(\varphi), ...\}$ and $F(d') = \{s'_1(\varphi), s'_2(\varphi), ...\}$, respectively. The Euclidean distance between them is calculated according to the formula.

This section uses features for Chinese-English feature correspondence, filtering paragraphs that do not meet the correspondence requirements and significantly narrowing the scope of plagiarism results. The Dice coefficient is then used to calculate similarity between text d and text d', as shown in Equation (11):

$$\mathrm{sim}(d, d') = \frac{2 \times |F(d) \cap F(d')|}{|F(d)| + |F(d')|}$$

Thus, the similarity between the two texts can be obtained.

---

4 Experiments and Validation

The experimental environment is configured as follows:

- **Experimental Platform:** Windows 7 (64-bit)
- **Processor:** Pentium (R) Dual-Core @ 2.50 GHz
- **Memory:** 4.00 GB
- **Experimental Environment:** MyEclipse, WinPython-64bit-2.7.10.3
- **Development Languages:** Java, Python

Experimental Data: The experimental data is divided into training and test datasets.

- **Training Dataset:** 3,500 articles from the Computer Science discipline in Springer were automatically translated into Chinese as positive samples. Negative samples were sourced from 2,800 Chinese articles in the China Academic Journal Network Publishing Database under the Computer Software and Computer Applications category, including prestigious journals such as *Journal of Computer Science and Technology* and *Journal of Software*.
- **Test Dataset:** 100 English articles from Springer with their Chinese translations and 50 Chinese articles from CNKI.

4.1 First Filtering

After text preprocessing, features are extracted and low-frequency features are removed. Information entropy is calculated for eligible features to determine each feature term' s stability. Next, the three weight parameters k_1, k_2, and k_3 must be determined. Manual tuning is too complex and clearly impractical. Through manual ranking and Algorithm 1, the optimal parameters are selected: $k_1 = 0.04$, $k_2 = 0.78$, $k_3 = 0.13$. The comparison results are shown in Figure 4 [Figure 4: see original paper].

After obtaining feature weights, articles in the training set are represented by features, and SVM is used to train the classifier, yielding the classification model.

Few studies exist on plagiarism classification based on translation features. This paper compares three evaluation metrics: (1) results from training after feature selection and weight assignment, (2) results from training after feature selection but before weight assignment, and (3) results from feature training provided in literature [18]. Using the paper' s training and test datasets, both closed and open tests were conducted, with results compared in Figures 5 [Figure 5: see original paper] and 6 [Figure 6: see original paper].

The figures show that in closed testing, the proposed method significantly outperforms literature [18] in all metrics except recall on non-plagiarized texts, where it ties, with superior comprehensive F-values. In open testing, this advantage is more pronounced, with all metrics leading. Therefore, the proposed method substantially improves feature selection accuracy for cross-language plagiarism, proving the effectiveness of the feature selection approach.

4.2 Second Filtering

For multiple English paragraphs corresponding to one Chinese paragraph, all non-compliant paragraphs are filtered out. After comparing 1,000 paragraph pairs, 749 retained only one suspicious paragraph after two rounds of filtering, with 736 accurately matching the plagiarized paragraph and only 13 mismatches. In the remaining 251 paragraph candidate sets, 24 had no matching suspicious

paragraphs, while 227 had multiple matching suspicious paragraphs. Figure 7 [Figure 7: see original paper] illustrates these results, showing that accuracy has reached 74% at this stage. For the 227 paragraphs matching multiple results, WordNet is employed to identify the specific plagiarized paragraph.

Statistics show that among the 227 validated paragraphs, 220 achieved accurate plagiarism correspondence, with only 7 screening errors. These errors occurred because the correct paragraph did not obtain the maximum similarity in WordNet similarity calculation. However, the plagiarized paragraph existed among the suspicious paragraphs after filtering, indirectly demonstrating result validity.

This paper applies two rounds of filtering based on feature correspondence followed by WordNet-based cross-language similarity detection, and compares it with literature [18]' s direct WordNet-based cross-language similarity detection. The precision, recall, and F-value comparison is shown in Figure 8 [Figure 8: see original paper].

The results show that both precision and recall are improved through the proposed method. The improvement stems from two rounds of filtering that eliminate paragraphs with similar semantics but significant differences in translation and structural features, retaining only paragraphs with small differences in translation and structural features but not necessarily close semantics. This significantly enhances precision and validates the theoretical effectiveness of the proposed approach.

---

5 Conclusion

The proposed method bridges inconsistencies between language and grammar, approaching plagiarism detection from a novel perspective. However, as previously noted, cross-language plagiarism detection is still in its infancy with many shortcomings requiring continuous improvement. First, corpus quality directly affects classification training results, necessitating future efforts in building high-quality corpora. Second, feature construction requires further improvement and mining, with automatic mining of translation features being a future research focus. Finally, efficiency issues require greater attention, particularly when facing large-scale datasets, which represents another key area for future research.

References

[1] Zhang Gege, Sun Meiying. Ethical issues within the scientific community: taking institutions of higher learning and scientific research institutions as examples [J]. JuanZong, 2015 (6): 622-625.

[2] Brassil J T, Low S, Maxemchuk N F, et al. Electronic marking and identification techniques to discourage document copying [J]. IEEE Journal on Selected

Areas in Communications, 1995, 13 (8): 1495-1504.

[3] Kang Cunhui. Detection of academic misconduct from the perspective of Moral Governance [J]. Journal of Wuhan Textile University, 2015 (2): 74-76.

[4] Zou Du, Chen Yuqing, Zhang Ling. Method of plagiarism detection based on semantic matching [J]. Journal of South China University of Technology: Natural Science Edition, 2013, 41 (7): 131-136.

[5] Zhang Wei. Research on Chinese text copy detection based on n-gram [D]. Changsha: Hunan University, 2014.

[6] Xia Zhiming, Liu Xin. A Chinese text similarity algorithm based on semantics [J]. Computer and Modernization, 2015 (4): 6-9.

[7] Zhang Xiankun, Zhang Qian. Research on ontology based comprehensive weighted case similarity algorithm [J]. Application Research of Computers, 2017 (2): 422-425.

[8] Xie Songshan, Tang Yan. Chinese text plagiarism detection algorithm based on left word frequency vector space model [J]. Journal of Southwest University: Natural Science Edition, 2015, 37 (5): 158-161.

[9] Zhu Qunyan. Research on cross language information retrieval based on comparable corpus [D]. Wuhan: Huazhong Normal University, 2015.

[10] Franco-Salvador M, Gupta P, Rosso P. Knowledge graphs as context models: improving the detection of cross-language plagiarism with Paraphrasing [C]// Proc of Bridging Between Information Retrieval and Databases. Berlin: Springer, 2014: 227-236.

[11] Franco-Salvador M, Rosso P, Montes-Y-Gómez, et al. A systematic study of knowledge graph analysis for cross-language plagiarism detection [J]. Information Processing & Management, 2016, 52 (4): 550–570.

[12] Peng Zhe. Research on cross language text correlation detection technology [D]. Changsha: Central South University, 2014.

[13] Liu Jiao, Cui Rongyi, Zhao Yahui, et al. An analysis method of cross-lingual literature similarity [J]. Journal of Yanbian University: Natural Science, 2016, 42 (2): 151-155.

[14] Zhang Jing. Classification feature selection algorithm for high-dimensional small sample data [D]. Hefei: Hefei Polytechnic University, 2014.

[15] Mcnamee P, Mayfield J. Character n-gram tokenization for European language text retrieval [J]. Information Retrieval, 2014, 7 (1-2): 73-97.

[16] Pu Xiaoyan. An analysis of the titles, English translations and definitions of "English Chinese" [J]. Journal of Nanchang College of Education, 2015 (12): 163-180.

[17] Nitto E D, Matthews P, Petcu D, et al. Model-driven development and operation of multi-cloud applications [M]. Berlin: Springer International Publishing, 2017.

[18] Yang Qianru. Research on cross language plagiarism detection technology based on fingerprint fusion [D]. Harbin: Harbin Engineering University, 2016.

[19] Franco-Salvador M, Gupta P, Rosso P. Cross-language plagiarism detection using a multilingual semantic network [C]// Proc of European Conference on Advances in Information Retrieval. Berlin: Springer, 2013: 710-713.

[20] Luo Yuansheng, Wang Mingwen, Le Zhongjian, et al. Bilingual topic correlation model in cross language information retrieval [J]. Journal of Chinese Computer Systems, 2013, 34 (12): 2758-2763.

[21] Narducci F, Palmonari M, Semeraro G. Cross-lingual link discovery with TR-ESA [J]. Information Sciences, 2017, 394-395: 68-87.

[22] Gamallo P, Pereira-Fariña M. Compositional semantics using feature-based models from wordNet [C]// Proc of Workshop on Sense. 2017.

Note: Figure translations are in progress. See original paper for figures.