

Postprint: Gender Identification of Chinese Weibo Authors Based on Dependency Relations

Authors: Qi Ruihua

Date: 2017-11-08T00:00:00+00:00

Abstract

Abstract

Purpose: To address the characteristics of short length and sparsity of traditional stylistic feature sets in online texts, this study investigates the application of dependency relations in Chinese microblog author gender identification.

Method: Publicly available Tencent microblogs are selected as experimental corpus. Dependency relation features are extracted and compared against lexical features, structural features, function word features, part-of-speech tagging features, and microblog features from existing literature through comparative experiments.

Results: Comparative experiments utilizing Support Vector Machine, Naive Bayes, k-Nearest Neighbors, and Decision Tree algorithms demonstrate that the proposed method achieves the highest accuracy, recall, and F-Measure for the Chinese microblog author gender identification task.

Limitations: The effectiveness of dependency relations in microblog author gender identification requires further validation on large-scale corpora.

Conclusion: The proposed model mitigates the sparsity of short text feature sets and, compared with other control feature sets, more effectively identifies author gender.

Full Text

Identifying Chinese Microblog Author Gender Based on Dependency Relations

Qi Ruihua

School of Software, Dalian University of Foreign Languages, Dalian 116044, China

Abstract

[Objective] This study investigates the application of dependency relations in Chinese microblog author gender identification, addressing the challenges posed by the short length and sparse traditional stylistic feature sets of online texts. **[Methods]** We selected public posts from Tencent Weibo as experimental corpus and extracted dependency relation features to conduct comparative experiments against existing lexical, structural, function word, part-of-speech tagging, and microblog-specific features reported in the literature. **[Results]** Controlled experiments using Support Vector Machine, Naive Bayes, k-Nearest Neighbors, and Decision Tree algorithms confirmed that our method achieves the highest precision, recall, and F-Measure for Chinese microblog author gender identification. **[Limitations]** The effectiveness of dependency relations in microblog author gender identification requires further validation on larger-scale corpora. **[Conclusions]** The proposed model effectively avoids the sparsity issue inherent in short text feature sets and demonstrates superior performance in author gender identification compared to existing feature sets.

Keywords: Dependency Relations, Chinese Microblog, Gender Identification

1. Introduction

With the rapid proliferation of online applications, massive amounts of web text have emerged, making author attribute analysis a hot topic in fields such as marketing and cyber forensics. Twitter alone generates over 500 million new messages daily, while user identity theft has become increasingly prevalent, with more than 32 million Twitter users' login credentials compromised in 2016 [1], and such cases continuing to rise annually. The surge in social media users and information volume has underscored the urgency of author attribute research.

Author gender analysis represents a primary task in identity attribute research. Analyzing author gender in web texts enables businesses to conduct targeted marketing, thereby improving the efficiency of personalized recommendations and market expansion. Gender analysis also helps identify the sources of anonymous false information and misinformation, preventing severe negative impacts on socio-economic order and public security.

Microblogs have become an important focus for author gender analysis. In Q1 2016 alone, Sina Weibo's monthly active users reached 261 million, representing a 32% year-over-year increase [2]. Microblog author gender identification has attracted significant research attention both domestically and internationally, including studies that infer author gender from Twitter user information and tweet content [3], or construct ensemble classifiers for Chinese microblog author gender using usernames and post text [4]. Existing methods, however, rely heavily on username information and fail to account for cases where authors deliberately conceal their identities.

To address this limitation, this paper proposes an author gender identification method that requires no microblog user information. By extracting dependency relation features from microblog text, we construct a stylistic feature model for microblog author gender identification and validate the effectiveness of dependency relation features through comparative experiments against established feature sets.

2. Related Research on Author Gender Identification

Research on web text author gender analysis has primarily focused on English corpora, including online reviews, BBS, and blog posts. Representative studies include Schler et al. [5], who analyzed tens of thousands of English blog posts totaling nearly 300 million words, confirming significant differences in writing style and content between males and females. Argamon et al. [6-7] combined linguistic features such as pronouns, determiners, prepositions, and content features with Bayesian Multinomial Regression for blog author gender analysis, achieving approximately 70% accuracy. Mikros et al. [8] utilized blog corpora from 20 Greek authors to construct stylistic feature sets including word length statistics, lexical richness, most frequent words, and character N-grams, achieving over 80% gender identification accuracy using Support Vector Machines. Rangel et al. [9] demonstrated that stylistic features such as word frequency, punctuation, part-of-speech tagging, and English and Spanish sentiment words aid in identifying anonymous authors' gender, achieving 57% accuracy on the PAN-AP-133 dataset using SVM. These studies generally used longer texts than microblogs, with feature sets ranging from hundreds to thousands of dimensions, resulting in obvious feature sparsity.

Burger et al. [3] extracted character 1-5grams and word 1-2grams from Twitter users' nicknames, account names, personal descriptions, and tweet content, achieving up to 92% accuracy in gender identification, though using only tweet text features yielded only approximately 75% accuracy.

For Chinese corpora, Tang et al. [10] extracted gender-specific descriptive words and address terms from Chinese novels, noting that the former better indicates gender, and achieved 73.2% accuracy in name gender identification using combined features, though this method was not validated on short texts. Huang et al. [11] proposed a rough set-based microblog user gender identification algorithm using term feature vectors, with an improved term frequency weighting mechanism that reduced document zero-similarity phenomena, but did not address how to determine tolerance thresholds. Bai [12] selected texts from Tianya forum's automotive and stock sections, obtained feature words through CFS and BestFirst algorithms, and achieved 70%-80% accuracy using Naive Bayes and SVM, though accuracy depended heavily on text length. While content-based feature words improved gender identification accuracy, they compromised cross-topic applicability. Wang et al. [4] used username 1-2grams and first-character features combined with microblog text 1-2gram features to construct a Bayesian classification fusion algorithm, achieving approximately 90% accuracy, though

using only microblog text features yielded only about 74% accuracy.

Deep syntactic dependency analysis can extract topic-independent abstract syntactic structure information, thereby revealing implicit writing habits [13], and has recently been applied to author style analysis. For example, Hollingsworth [14] used DepWords encoding instead of traditional syntactic dependencies and utilized their statistical features to identify authors of English detective novels. Zhang et al. [15] extracted features including structural features, function words, POS, common words, and dependency relations, with comparative experiments on 21 English works and Reuters corpora demonstrating that dependency relations help improve author identification efficiency. However, the effect of dependency relations on author gender identification remains unexplored.

Based on analysis of existing research and microblog text characteristics, this paper proposes a novel dependency relation-based stylistic feature model for author gender identification.

3. Author Gender Stylistic Feature Model

Let the author gender set be $\{A_{Female}, Male\}$. Given a training set, the task of automatic author gender identification is to learn an author gender feature model from the training set and assign the most probable gender t_G ($t_G \in A$) to an anonymous text t . To accomplish this task, unstructured text must first be mapped into a stylistic feature vector space to extract an author gender stylistic feature set. This feature set should possess descriptive power for distinguishing author gender, and its feature values should be readily obtainable.

3.1 Dependency Relations Dependency relations, proposed by French linguist Tesnière et al. [16], constitute a theoretical framework for describing syntactic structure based on subordinate and dominant relationships between words. They have been widely applied in text mining, multilingual processing, semantic annotation, and information retrieval. A dependency relation consists of dependency pairs between a sentence's head word and dependent words. Let w_i be the i -th word in a sentence. After extracting dependency relations, a sentence can be represented as $R(w_i, w_j)$, where each dependency relation $R_i \in R(w_i, w_j)$ forms a directed arc from the governing word w_{2_i} to the dependent word w_{1_i} , with $w_i, w_j \in V$ and R being the set of all dependency relation types. The formal axioms of dependency relations include [17]: a sentence has only one independent component; all other components directly depend on some component in the sentence; no component can depend on more than two components; if component X directly depends on component Y, and component Z is positioned between X and Y in the sentence, then Z depends on X or Y, or on some component between X and Y.

Dependency relations offer three advantages as stylistic features for author gender identification: they have simple storage structures, good computability, and strong adaptability to big data web texts and cross-language environments; de-

dependency parsing emphasizes dominant-dependent and modifier-modified relationships between sentence components, with order-independent characteristics that facilitate analysis of flexible web text sentence patterns; additionally, dependency relations extract abstract syntactic structure information with content independence.

This paper adopts the 22 Chinese dependency relation types defined in FudanNLP [18] as the author gender identification feature set: $F_{dependency} = \{\text{associative, subject, punctuation, interrogative-serial, complement, voice, de-attributive, prepositional-object}\}$. An example of sentence component dependency relations is shown in Figure 1 [Figure 1: see original paper].

3.2 Existing Stylistic Features This paper incorporates major stylistic features proposed in existing literature for comparative experiments, including lexical features, structural features, function word features, part-of-speech tagging features, and microblog features.

Lexical features include statistical characteristics and frequencies of words, such as word length, lexical richness, word frequency, word N-grams, and special vocabulary. Lexical feature extraction largely depends on corpus length and is therefore typically not used alone. Considering the short length of microblogs, to avoid lexical feature sparsity, this paper selects content-independent numerals, high-frequency words, temporal words, and date words based on the 2015 Green Book on China's Language Situation published by the National Language Resources Monitoring and Research Center [19] as lexical features in comparative experiments.

Structural features include text organization and layout-related features such as punctuation marks, paragraph count, paragraph length, and average sentence length, which are particularly effective for short texts like emails, blogs, or microblogs. Following reference [15], this paper selects sentence count, character count, and occurrences of colons, semicolons, thousand/percent signs, unit symbols, periods, left/right quotation marks, left/right parentheses, commas, exclamation marks, ellipsis (single), ellipsis (double), dashes, spaces, question marks, and enumeration commas as structural features.

Function word features refer to words that lack independent lexical meaning and only express grammatical meaning or function, exhibiting topic independence. In modern Chinese, function words are also called empty words. Function words appear frequently in small quantities and have proven effective as stylistic features [20]. Chinese function words play a more important grammatical role, bearing grammatical meanings expressed by content word inflections in Western languages. For comparative experiments, this paper selects the combined set of Chinese function words from references [20-21] as function word features.

Part-of-speech tagging features involve word type annotation based on word form or syntactic behavior, typically without involving specific word meanings and exhibiting topic independence. The ICTCLAS [22] Chinese POS tagset from

the Institute of Computing Technology, Chinese Academy of Sciences includes 22 first-level tags, 66 second-level tags, and 22 first-level POS tags' occurrences per thousand words.

Microblog features include text layout formats specific to microblog text, such as topic citations, username references, and image hyperlink usage. Following reference [23], this paper counts the occurrences of images, URLs, # symbols, @ symbols, email addresses, and emoticons in microblogs.

4. Experiments and Analysis

4.1 Data Preparation We selected public posts from Tencent Weibo as experimental corpus, collecting 6,530 microblog posts from verified public figures over a 10-month period in 2012, including 5,496 posts by male authors and 1,034 by female authors. The longest microblog text contained 284 characters, the shortest 5 characters, with an average text length of 73 characters; 65% of samples contained fewer than 100 characters.

The experiments used ICTCLAS 2015 [22] for Chinese word segmentation and POS tagging, FudanNLP 1.5 [18] for dependency syntactic analysis, and Weka 3.7.9 [24] as the classification algorithm environment. Ten-fold cross-validation was performed in comparative experiments, with precision, recall, and F-Measure used to evaluate model performance.

The Chinese dependency relation feature set was compared against major stylistic features from literature [20-21, 23], including lexical features, structural features, function word features, POS tagging features, and microblog features. The feature sets and their dimensions are illustrated in Figure 2 [Figure 2: see original paper].

4.2 Experiments and Analysis To validate the effectiveness of dependency relations in Chinese microblog author gender identification, comparative experiments were conducted using four classification algorithms: Support Vector Machine (LibSVM), Naive Bayes (NBC), k-Nearest Neighbors (IBK), and Decision Tree (C4.5). Experimental results are presented in Table 1, with the highest values for each algorithm highlighted in bold.

(1) Comparing feature set discrimination effectiveness: Overall, the dependency relation feature set achieved the highest precision, recall, and F-Measure values in Chinese microblog dataset experiments, with all three key metrics reaching over 99.7% for SVM, k-NN, and Decision Tree algorithms. Results confirm that the dependency relation feature set can mine deep syntactic features in microblog text expression across genders, demonstrating better adaptation to short texts compared to lexical, structural, function word, POS, and microblog features, thereby avoiding the impact of feature set sparsity on algorithm efficiency.

(2) Algorithm performance analysis: Overall, k-NN, SVM, and Decision

Tree C4.5 algorithms achieved higher weighted averages for precision, recall, and F-Measure, while Naive Bayes performed moderately. This occurs because Naive Bayes' independence assumption does not hold for most feature sets, whereas k-NN and SVM can adapt to noise in samples, and Decision Tree uses information gain ratio for feature selection, overcoming feature sparsity and noise interference in short texts.

(3) Naive Bayes performance: In Naive Bayes experiments, the dependency relation feature set performed worse than other feature sets because dependency relation features violate Naive Bayes' independence assumption.

(4) Key feature analysis: Figure 3 [Figure 3: see original paper] shows the decision tree constructed by the C4.5 algorithm for gender identification using microblog text dependency relation features. Key features are relatively concentrated, including associative relations, subject relations, tense relations, zhi-constructions, de-constructions, and serial verb constructions, which can serve as primary features for further investigation into gender-specific tendencies in syntactic structure selection.

This paper explores the application of deep syntactic analysis features in Chinese microblog author gender analysis. Experimental results demonstrate that compared to existing methods, the proposed dependency relation-based stylistic feature model effectively avoids short text feature set sparsity and more effectively identifies author gender. We found that associative relations, subject relations, tense relations, zhi-constructions, de-constructions, and serial verb constructions play key roles as decision nodes. Future research will validate these findings on larger-scale corpora.

References

- [1] Sina Technology. 32 Million Twitter Account Stolen [R/OL]. [2016-06-09]. <http://tech.sina.com.cn/i/2016-06-09/doc-ifszmaa1783949.shtml>.
- [2] Sina Technology. Micro-blog Monthly Active Users Increased to 261 Million [R/OL]. [2016-05-12]. <http://tech.sina.com.cn/i/2016-05-12/doc-ifsxenvm0294013.shtml>.
- [3] Burger J D, Henderson J, Kim G, et al. Discriminating Gender on Twitter[C]//Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011: 1301-1309.
- [4] Wang Jingjing, Li Shoushan, Huang Lei. User Gender Classification in Chinese Microblog [J]. Journal of Chinese Information Processing, 2014, 28(6): 150-155, 168.
- [5] Schler J, Koppel M, Argamon S, et al. Effects of Age and Gender on Blogging [C]//Proceedings of the 2006 Association for the Advance of Artificial Intelligence Spring Symposium: Computational Approaches to Analyzing Weblogs.

2006.

- [6] Argamon S, Koppel M, Pennebaker J W, et al. Automatically Profiling the Author of an Anonymous Text[J]. *Communications of the ACM*, 2009, 52(2): 119-123.
- [7] Argamon S, Koppel M. A Systemic Functional Approach to Automated Authorship Analysis[J]. *Journal of Law & Policy*, 2013, 12: 299-315.
- [8] Mikros G K, Perifanos K. Authorship Attribution in Greek Tweets Using Author' s Multilevel N-Gram Profiles[C]//*Proceedings of the 2013 Association for the Advance of Artificial Intelligence (AAAI) Spring Symposium: Analyzing Microtext*. 2013.
- [9] Rangel F, Rosso P. Use of Language and Author Profiling: Identification of Gender and Age[C]//*Proceedings of the 10th Workshop on Natural Language Processing and Cognitive Science*. 2013.
- [10] Tang Qin, Lin Hongfei. Research on Gender Recognition for Character in Text [J]. *Journal of Chinese Information Processing*, 2010, 24(2): 46-51.
- [11] Huang Faliang, Xiong Jinbo, Huang Tianqiang, et al. Gender Identification of Microblog Users Based on Rough Set[J]. *Journal of Computer Applications*, 2014, 34(8): 2209-2211.
- [12] Bai Lijuan. Gender Classification Based on Text Mining [D]. Harbin: Harbin Institute of Technology, 2011.
- [13] Qi Ruihua, Yang Deli, Guo Xu, et al. Blogger Identification Based on Multidimensional Stylistic Features[J]. *Journal of the China Society for Scientific and Technical Information*, 2015, 34(6): 628-634.
- [14] Hollingsworth C. Using Dependency-based Annotations for Authorship Identification[M]. *Text, Speech and Dialogue*. Springer Berlin Heidelberg, 2012: 314-319.
- [15] Zhang C, Wu X, Niu Z, et al. Authorship Identification from Unstructured Texts[J]. *Knowledge-Based Systems*, 2014, 66: 99-110.
- [16] Tesnière L, Osborne T, Kahane S. *Elements of Structural Syntax*[M]. John Benjamins Publishing Company, 2015.
- [17] Robinson J J. Dependency Structures and Transformational Rules[J]. *Language*, 1970, 46(2): 259-285.
- [18] Fudan Natural Language Processing Group. FudanNLP[EB/OL]. [2016-01-01]. <http://nlp.fudan.edu.cn/software/>.
- [19] National Language Resources Monitoring and Research Center. Chinese Language Situation over the Years [R/OL]. [2015-01-01]. <http://cnlr.blcu.edu.cn/col/col8765/index.html>.
- [20] Zheng R, Li J, Chen H, et al. A Framework for Authorship Identification of Online Messages: Writing-style Features and Classification Techniques[J]. *Jour-*

nal of the American Society for Information Science and Technology, 2006, 57(3): 378-393.

[21] Yu B. Function Words for Chinese Authorship Attribution[C]//Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2012.

[22] ICTCLAS 2015[EB/OL]. [2015-01-01]. <http://ictclas.nlpir.org/downloads>.

[23] Silva R S, Laboreiro G, Sarmiento L, et al. ‘twazn me!!!; (‘ Automatic Authorship Analysis of Micro-blogging Messages[M]. Natural Language Processing and Information Systems. Berlin Heidelberg: Springer, 2011: 161-168.

[24] Machine Learning Group at the University of Waikato. WEKA[EB/OL]. [2015-01-01]. <http://www.cs.waikato.ac.nz/ml/weka/downloading.html>.

Conflict of Interest Statement

The author declares no conflict of interest.

Supporting Data

Supporting data is self-archived by the author, E-mail: rhqi@dluf.edu.cn.

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

Source: ChinaXiv – Machine translation. Verify with original.