View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 474 – 478

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

Automatic Genre Classification via N-grams of Part-of-Speech Tags

Xiaoyan Tang[a,]*, Jing Cao[a]

[a]*Zhongnan University of Economics and Law, Wuhan 430073, P.R.China*

Abstract

Recurring sequences of words have long been considered as a signifier of different genres and registers by corpus linguists. The previous research mainly focused on lexical n-grams. Nevertheless, n-grams of other linguistic features, such as part-of-speech, have been less studied. The current study is expected to examine whether n-grams of part-of-speech tags extracted from a large corpus can be a discriminator of different genres. The results show that a strong correlation exists between the information about n-grams of part-of-speech tags and the genre of the text.

© 2015 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

Keywords: Automatic Genre Classification; BNC Baby; N-grams; Naïve Bayes Classifier; Multinomial Naïve Bayes Classifier; Part-of-Speech

1. Introduction

Recurring sequences of words have long been considered as a signifier of different genres and registers by corpus linguists (e.g. Biber & Barbieri, 2007; Biber *et al*., 2004; Chen & Baker, 2010; Cortes, 2004), since Biber *et al*. (1999) observed that the internal linguistic features of lexical n-grams are different in conversation and academic prose. Biber *et al*. (2004) analyzed the frequencies, structural types and functional categories of n-grams and their distributions in university teaching and textbooks, and was extended by Biber & Barbieri (2007) to a wider range of spoken and written university registers. Cortes (2004) made a comparison between publications and student writings in history and biology. Chen & Baker (2010) did structural and functional analysis of n-grams in corpora of L1 and L2 academic writing. Besides, Gries (2010a, 2010b, 2011) explored the n-gram frequencies among various registers

* Corresponding author.
E-mail address: 2011txy@gmail.com

1877-0428 © 2015 Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.468

with several advanced quantitative methods. The previous research mainly focused on lexical n-grams. Nevertheless, n-grams of other linguistic features, such as part-of-speech, have been much less studied (except Santini, 2004). Santini (2004) presented genre classification experiments using unigrams, bigrams and trigrams obtained from BNC, and trigrams gained the best performance.

The current study is expected to further examine whether n-grams of part-of-speech tags extracted from a large corpus can be a discriminator of different genres. BNC Baby, a genre balanced sub-corpus of BNC, is employed as the resource of part-of-speech n-grams and genres. There are two tag sets (CLAWS5 tag set and simplified part-of-speech tag set), five lengths of n-grams (n=1, 2, 3, 4, 5), and two classifiers (Naïve Bayes Classifier and Multinomial Naïve Bayes Classifier) used in the experiments. The results show that a strong correlation exists between the information about n-grams of part-of-speech tags and the genre of the text.

The remaining paper consists of the following sections. The second section introduces the source of data, the n-gram data preparation and classifiers used in this study. The third section outlines the experiment results. The fourth section discusses the findings and future research.

2. Methodology

In this section, the methodology of the current study is introduced. In the first and second subsections, the source of data used to generate n-grams and the n-gram data preparation process are presented respectively. In the third subsection, two classifiers and a machine learning workbench utilized in the current study are proposed, including how the results are achieved and evaluated.

2.1. Corpus and Tag sets

BNC Baby, a genre balanced sub-corpus of BNC, is employed as the resource of part-of-speech n-grams and genres. It consists of four one-million-word genre-based subsets (i.e. academic, fiction, newspaper and conversation) and is tagged with both CLAWS5[1] tag set (hereinafter, C5), which has over 60 tags and the simplified part-of-speech tag set (hereinafter, s-POS) which has 10 tags as listed in the table 1. Since Zipf's Law is "true for the frequency of occurrence of n-grams" (Cavnar & Trenkle, 1994), each text is used to generate the same quantity of n-grams.

Table 1. Simplified part-of-speech tag set

Tag	Part-of-speech
ADJ	adjective
ADV	adverb
ART	article
CONJ	conjunction
INTERJ	interjection
PREP	preposition
PRON	pronoun
SUBST	substantive
UNC	unclassified, uncertain, or non-lexical word
VERB	verb

[1] See http://ucrel.lancs.ac.uk/claws5tags.html for the description of entire tag set.

2.2. Data preparation

In order to classify texts according to their part-of-speech features, the two tag sets and BNC Baby needs to be processed. First, n-grams of different lengths are generated from the tag set. Any n-gram set is the set of permutations of any n tags in the tag set. For example, the unigram set is the set of all tags in the tag set, and the bigram set are the set of permutations of any two tags in the tag set. Second, the frequency information of every n-gram generated from tag set in every text is calculated from BNC Baby. For each combination of tag sets and different lengths of n-grams (e.g. trigrams of tags in C5), a frequency lists is constructed. Third, after being sorted, the 300 most frequent tags and their frequency information in every frequency list are left and others are abandoned to reduce the cost of computation when conducting the classification experiments. All frequency lists will be taken as inputs when the classification experiments are conducted.

2.3. Classifiers

The Naïve Bayes Classifier and Multinomial Naïve Bayes Classifier in Weka (Hall *et al*. 2009) are used for automatic genre classification. These two classifier are both probabilistic classifiers that based on Bayes' theorem and assume independence among features (in this study part-of-speech tags). But the former one takes advantage of the knowledge of existence of different features, while the latter one makes use of the frequency information of different features. Weka is a popular free machine learning workbench that enables the user to implement various classification algorithms including the two applied in the current study.

Stratified 10-fold cross validation is used and repeated ten times to calculate the results of two classifiers on two tag sets and part-of-speech n-grams of five different lengths. And the results are evaluated by average weighted F-score.

3. Experiment Results

As mentioned in the last section, two classification algorithms, Naïve Bayes and Multinomial Naïve Bayes, are implemented in Weka in the current study. Weka takes a frequency list constructed from last section as the input, and presents an average weighted F-scores as the output. Since there are two classifiers, two tag sets and five sets of n-grams for each tag set, there are twenty experiments, thus twenty F-scores in total. Table 2 below synthesizes all experiment results.

Table 2. Average weighted F-scores

N-grams	C5		s-POS	
	Naïve Bayes	Multinomial Naïve Bayes	Naïve Bayes	Multinomial Naïve Bayes
1-gram	0.913	0.904	0.888	0.888
2-grams	0.946	0.931	0.899	0.899
3-grams	0.962	0.921	0.908	0.925
4-grams	0.951	0.921	0.935	0.942
5-grams	0.956	0.926	0.946	0.947

According to the table above, several findings could be summarized:

- All the weighted average F-measure obtained from this study range from 0.888 to 0.962, indicating pretty strong correlation between the occurrences or frequencies of part-of-speech n-grams and genre. In general, the results also echo the previous studies (e.g. Santini, 2004) in that the bigger the n is, the better results will be achieved.

- It can also be observed that when the n-grams are obtained from C5, the Naïve Bayes Classifier always performs better than Multinomial Naïve Bayes Classifier. However, if the n-grams are extracted from s-POS tagging, the Multinomial Naïve Bayes Classifier tends to perform better in five out of the twenty experiments.

In addition, interesting results merged when we take a close look at the F-measure of the individual genres. Of all the twenty experiments, twelve have the best prediction for conversation, seven for fiction, one for academic, and none for newspaper. Therefore, the findings invite further research with larger corpus data and a wider range of genres as well. The results are listed below.

Table 3. Best prediction in terms of individual genres

N-grams	C5		s-POS	
	Naïve Bayes	Multinomial Naïve Bayes	Naïve Bayes	Multinomial Naïve Bayes
1-gram	Fiction	Conversation	Fiction	Conversation
2-grams	Fiction	Conversation	Fiction	Conversation
3-grams	Academic	Conversation	Fiction	Conversation
4-grams	Conversation	Conversation	Fiction	Conversation
5-grams	Conversation	Conversation	Fiction	Conversation

4. Discussion and Conclusion

This study shows that n-grams of part-of-speech tags extracted from a large corpus can be a discriminator of different genres since a strong correlation exists between the information about n-grams of part-of-speech tags and the genre of the text. To be more specific, the correlation is stronger when the length of n-grams is longer, which is in accordance with both previous studies and the intuition since longer n-grams carry more syntactic knowledge about the text. And considering C5 is richer than s-POS, that when the length of n-grams increases, Naïve Bayes Classifier always performs better than Multinomial Naïve Bayes Classifier on C5 but worse on s-POS may results from the balance between the information included in the features and information being used when applied to the classifier.

Nevertheless, the current study has many limitations, such as the size of the corpus is relatively small, and the genre system used in classification, which has only four genres, is very simple. Future research should employ larger corpus, for example the entire BNC, and more fine-grained genre classification criteria.

References

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for specific purposes*, 26(3), 263-286.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. *Applied linguistics*, 25(3), 371-405.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., & Quirk, R. (1999). *Longman grammar of spoken and written English*. London/New York.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology*, 14(2), 30-49.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for specific purposes*, 23(4), 397-423.

Gries, S. T. (2010a). Bigrams in registers, domains, and varieties: a bigram gravity approach to the homogeneity of corpora. In *Proceedings of Corpus Linguistics 2009*, University of Liverpool.

Gries, S. T., & Mukherjee, J. (2010b). Lexical gravity across varieties of English: an ICE-based study of n-grams in Asian Englishes. *International Journal of Corpus Linguistics*, 15(4), 520-548.

Gries, S. T., Newman, J, & Shaoul, C. (2011). N-grams and the clustering of registers. *Empirical Language Research*, 5.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.

Santini, M. (2004). A shallow approach to syntactic feature extraction for genre classification. In *proceedings of the 7th Annual Colloquium for the UK Special Interest Group for Computational Linguistics* (pp. 6-7). Birmingham, UK.

The BNC Baby, version 2. 2005. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/