EPiC Series in Language and Linguistics

Volume 4, 2019, Pages 107–112

Proceedings of Third Workshop "Computational linguistics and language science"

Verb Constructions As A Feature Of Genre Classification

N.N.Bujlova[1]

[1] National Research University Higher School of Economics, Moscow, Russia

nbujlova@hse.ru

Abstract

In our research we studied the dependency structure of the text genre (love stories, detective stories, science fiction and fantasy). The novel characteristics (such as syntactic attributes as verb constructions and construction of a specific cumulative threshold) which can be additional machine learning parameters were identified. We have conducted an experiment with novel features and showed that these characteristics can be useful for closely related genre recognition.

Introduction

The text genre is one of the key concepts in literary studies and is useful both in theoretical and practical applications, for example, in the development of electronic libraries and corpora. Plenty of papers have studied the structure of various genres.

The existing methods of genre classification have good performance in classifying texts which belong to different registers [1]. Among the methods of genre determination, machine learning is currently the most popular and such methods as Naïve Bayes classifier, support vector machine, decision tree, random forest are extensively applied. These methods can work with minimally annotated texts (with given metatags only) as well as with the processed data. For instance, in the discrimination of scientific, news and fictional texts part-of-speech (POS) [2], [3], [4] characteristics are used; morphological [5], syntactic [6] and statistical [7] features can be applied for readability classification. There are many tools of genre classification designed for English, but for other languages the diversity of application is limited [8], [9], [10].

The particular task is the classification of closely related genres. Traditional methods (based on statistical [11] and discursive [12] characteristics, POS-histograms [13], etc), to the best of our knowledge, are not used to separate the closely related genres [14], at least with regard to the genre classification within Russian prose. This issue can potentially be solved by using not only the morphological features of the text, but also syntactic ones, such as verb constructions [15].

In our study, we investigate the relationship between the argument-predicate structure and text genre. We examine both verb frequencies and differences in verb constructions across four genres of

G. Wohlgenannt, R. von Waldenfels, S. Toldova, E. Rakhilina, D. Paperno, O. Lyashevskaya, N. Loukachevitch, S.O. Kuznetsov, O. Kultepina, D. Ilvovsky, B. Galitsky, E. Artemova and E. Bolshakova (eds.), CLLS 2018 (EPiC Series in Language and Linguistics, vol. 4), pp. 107–112

popular prose - love stories, detective stories, science fiction and fantasy, which are quite similar to each other and pose a challenge for existing classical methods of defining the text genre. We hypothesize that the use of additional features can improve the discrimination performance.

Dataset and preprocessing

Data were collected from Russian web library Librusek and Moshkov's library. We extracted four subcorpora: love stories (Size of corpus 7 400 231), detectives (14 313 177), science fiction (16 228 321) and fantasy (6 245 659).

The texts were preprocessed using the UDPipe tool [16] which allows one to create corpora with syntactic dependencies annotation.

Table 1. General Properties of Four Samples

	Love Stories	Detective Stories	Science fiction	Fantasy
Average length of phrase	14	13	14	15
Average length of word	5	5	5	7
% noun	26	25	25	25
% adjective	6	6	6	6
% verb	19	19	20	19
% adverb	6	6	6	6

Table 1 demonstrates the similarity of fictional genres. The general quantitative properties of four samples such as Average length of phrase, Average length of word and the ratio of the main parts of speech are almost identical across all genres, so we can suggest that the usage of simple methods in the task of genre classification is doubtful. For these reasons we introduce two syntactic features - verb construction and construction of a specific cumulative threshold - as classifying features.

We define verb construction as the set "head (verb) + dependents (subject, adjunct, clause etc.)". We have selected six types of dependents relations - four arguments (mandatory for grammaticality and semantic coherence of the sentence) and two adjuncts (non-obligatory elements):

Nsubj (subject) – ***Ты*** *зря прохаживаешься...*

Obj (object) – Даже как ***вас*** *зовут, и то не знаю!*

Ccomp (clausal complement) – Да еще убеждала, ***что ей нельзя делать аборт***

Xcomp (open clausal complement) – Хотел ее ***встретить*** *и заблудился.*

Obl (oblique nominal) – Ляпнул Леня и тут же пожалел ***об этом.***

Advcl (adverbial clause modifier) – И даже попугай сегодня не показывался, ***хоть и обожал скандалы.***

We take into account both types of dependences and their order in the sentence. This approach allows for estimation of core and peripheral constructions; the latter are significantly undervalued in practical studies. The word order in the sentence increases the number of construction combinations and takes into account inversions and others stylistically-flavored phrases.

UDPipe output file contains indexed phrases and for every word and punctuation mark in each sentence there are an index within the sentence, the lemmas, the part of speech, morphological and syntactic tags, the number of the head and the type of connection with it. The initial annotation allows to select the head and dependencies, however, postprocessing of the data is necessary for the obtaining of the whole constructions. To extract the constructions we introduced several additional entities:

UniqueID (unique index of the token in the text in the format “number of phrase_number of token”), UniqueHead (the head index in the format “number of phrase_number of head”) and HeadLemma (root infinitive). If the head had a selected type of dependency, it was extracted in the form of construction. The frequency of each construction has been calculated for each file, each verb and each subcorpus.

We found the correlation between the number of dependencies and occurrence of construction in text: the more complex constructions appear rarely (Spearman correlation coefficient -0.59, p-value 2.2e-14, Figure 1).

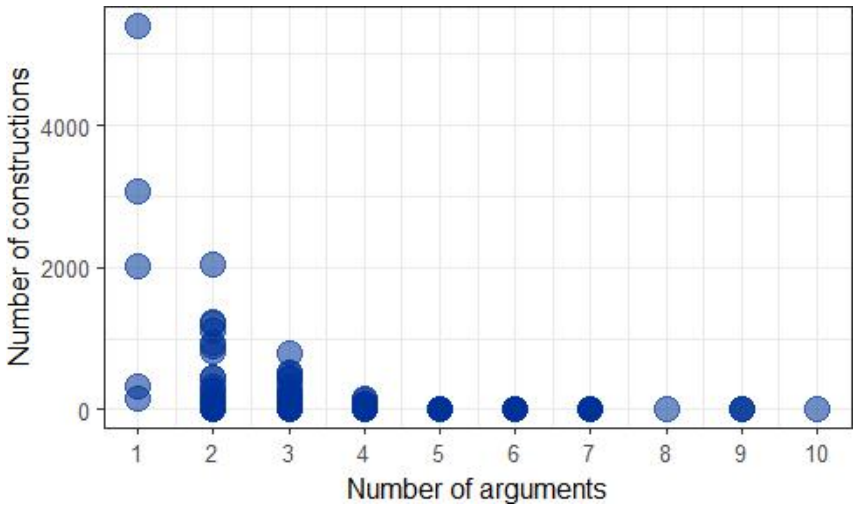

Figure 1**.** The relationship between the number of arguments and the frequency of the construction.

Types of verbs in corpora

For each genre we calculated frequency of occurrence of verb constructions and selected verb constructions with frequency more than 20. We compared the most frequent construction for every verb between genres which allowed us to identify syntactic markers for love stories, detective, sci-fi and fantasy (Table 2).

We found three types of verbs in our corpora. The verbs of the first type have similar constructions in different genres; the verbs of the second type have different frequencies in different genres, and the third type is characterized by taking different constructions in different genres. The most discriminative are constructions with the verbs of the third type, while the constructions with the verbs of the first type form a list of stop-structures similar to stop word lists.

Table 2. Selected Examples of the Verbs (second type, have different frequencies in different genres) and Their Constructions

	Love Stories	Detective Stories	Science fiction	Fantasy
бояться	Nsubj	xcomp	xcomp	nsubj
	nsubj;obl	nsubj obl	nsubj	xcomp
	nsubj;xcomp		obl	obl
замечать	nsubj;advcl	nsubj	nsubj;	nsubj
	nsubj	obj	obj;	obj nsubj;
	nsubj;obl	nsubj;advcl	nsubj; advcl;	advcl
помнить	nsubj	nsubj	nsubj	nsubj
	nsubj;obj	nsubj; advcl	nsubj;obj	nsubj;obj
	obl		nsubj;advcl	obj
проходить	nsubj	obl	nsubj	nsubj
	nsubj; obl	nsubj	obl	obl
	obl	nsubj;obl	nsubj;obl	nsubj;obl
лежать	nsubj;obl	obl	obl	obl
	obl	nsubj; obl	nsubj;obl	nsubj;obl
	obl;nsubj	obl;nsubj	obl;nsubj	obl;nsubj
продолжать	nsubj;xcomp	nsubj	xcomp	nsubj;xcomp
	xcomp	nsubj;xcomp	nsubj;xcomp	xcomp
	nsubj	obj	obj	nsubj

However, there are few of such markers per genre, and using the number of dependencies as a feature alone does not provide robust result.

We developed an additional metric – a construction of a cumulative specific threshold (CST), which determines the number of different constructions attached to the verb which cover N% occurrences in a particular genre. We applied a threshold of 50% in this study. Depending on the design number, on which the threshold was crossed, the verbs were assigned a rank. The majority of verbs had 1 to 4 constructions covering half of verb occurrences (Table 3).

Table 3. Verbs with Different Number of Constructions: Verb Ranks in Different Genres

Rank	Love Stories	Detective Stories	Science fiction	Fantasy
1	56	38	44	53
2	59	36	40	53
3	14	16	18	15
4	1	1	0	1

The selective manual inspection of more complex constructions showed that their origin is more likely associated with machine annotation errors than with unique features of head verbs.

Applying of machine learning for genre classification

We used several classifiers for genre classification. As input for machine learning we used four original corpora and their preprocessed version with syntactic annotation. The volume of training data

was 80% of corpus and the remaining data 20% were used as test data. We used simple classifiers because our purpose was to show the contribution of novel features in genre recognition. Python library "sklearn" was used to conduct machine learning experiment.

Naïve Bayes classifier with features from Table 1 (simplest text features - average length of phrase, average length of word and percentage of part of speech) yields the poorest result, so we used it as a baseline for further improvements (Table 4). Random forest based on initial corpus performed almost 10% better; the most pronounced effect was seen on love stories (0.75 compared to 0.5). After we added two syntactic features the results of the Random forest classifier increased up to 16% (Table 4).

Table 4. Results of Machine Learning Classification

	Naïve Bayes with simplest text features			Random forest with simplest text features			Random forest with syntactic features		
	P	R	F1	P	R	F1	P	R	F1
Love Stories	0.50	0.12	0.20	0.75	0.38	0.50	0.83	0.62	0.71
Detective Stories	0.59	0.91	0.71	0.62	0.89	0.73	0.77	0.91	0.83
Science fiction	0.83	0.71	0.77	0.86	0.67	0.75	0.67	0.86	0.75
Fantasy	0.75	0.86	0.80	0.67	0.86	0.75	0.8	0.57	0.67
avg / total	0.65	0.67	0.62	0.72	0.7	0.68	0.77	0.76	0.75

P – precision, R – recall, F1 – f1-score.

5 Conclusion

We evaluate the use of CST as a feature in machine learning for discrimination of popular literature genres. Since we have found the difference in the number and composition of constructions as well as in proportion of certain constructions across genres, CST appears to be a promising feature. We tested various machine learning methods using different sets of features. Naïve Bayes classifiers was used as baseline, which we compare with random forest. Future development implies experiments with different CST, other types of syntactic dependencies of the verb, the order of dependencies and different types of argument's tags.

References

1. Mangalova E.S., Agafonov E.D. O probleme vydeleniya informativnyh priznakov v zadache klassifikacii tekstovyh dokumentov. Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika. 1 (2013).
2. Karlgren J., Cutting D. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. Proceedings of the 15thConference on Computational Linguistics - Volume 2. pp 1071– 1075, Stroudsburg, PA, USA (1994).
3. Feldman S. et al. Part-of-speech histograms for genre classification of text. Acoustics, Speech and Signal Processing 2009. pp 4781– 4784. ICASSP (2009).
4. Braslavskij P. Morfologicheskij stroj funkcional'nyh stilej (na materiale dokumentov Internet). Izvestiya Ural'skogo gosudarstvennogo universiteta 21, 9-17 (2001).
5. Reynolds R., Eduard S., Detmar M. A VIEW of Russian: Visual Input Enhancement and adaptive feedback. Proceedings of the third workshop on NLP for computer-assisted language learning. pp 101 – 111. (2016).

6. Mason J.E., Shepherd M., Duffy J. An n-gram based approach to automatically identifying web page genre. HICSS'09. 42nd Hawaii International Conference. P. 1–10 (2009).
7. Baerman M., Brown D., Corbett G. G. The Syntax-Morphology Interface: A Study of Syncretism. Cambridge, Cambridge University Press. (2005).
8. Stamatatos E., Fakotakis N., Kokkinakis G. Automatic text categorization in terms of genre and author. Computational linguistics 26(4), 471–495, (2000).
9. Amasyali F. M., Banu D. Automatic Turkish Text Categorization in Terms of Author, Genre and Gender. Springer-Verlag Berlin Heidelberg. 778-792. 2006.
10. Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M., Al-Rajeh, A.: Automatic Arabic text classification. JADT'08. 77–83 (2008).
11. Radošević, D., Dobša, J., Mladenić, D., Novak, M., Stapić, Z.: Genre Document Classification Using Flexible Length Phrases. Information and Intelligent Systems. Fakultet organizacije i informatike, Varaždin, (2006).
12. Webber B. Genre distinctions for Discourse in the Penn TreeBank. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP. 674–682 (2009).
13. Kessler B., Nunberg G., Schutze H. Automatic Detection of Text Genre. CoRR (1997).
14. Borisov L.A., Orlov Ju.N., Osminin K.P. Identifikacija avtora teksta po raspredeleniju chastot bukvosochetanij. Prikladnaja informatika. 26(2), 95-108 (2013).
15. Apresyan YU.D. EHksperimental'noe issledovanie semantiki russkogo glagola. M.: Nauka (1967).
16. Straka M., Haji J., Strakov J. UDPipe: Trainable Pipeline for Processing CoNLLU Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. LREC, pp. 4290-4297. (2016)