Verb Constructions as a Feature of Genre Classification

Nadezhda N. Bujlova

National Research University Higher School of Economics,
School of Linguistics, Moscow
E-mail: nbujlova@hse.ru

Abstract

Our study addresses genre classification, a topical issue in the web corpora research. Web corpora are often criticized for the lack of information regarding the genre structure, which limits their applicability. The classification of closely related genres is a particular challenge. In our study we use syntactical features for the discrimination between four genres of fiction: love stories, detectives, science fiction and fantasy. We determined the lists of verb constructions which corresponds to each genre and introduced new metrics CSCT describing the number of major constructions for each verb. Both verb construction lists and CSCT as well as other text characteristics were used as features for the machine learning. We used three methods of machine learning: Naïve Bayes, Decision Trees and Random Forest. The use of syntactical features improved both precision and recall, which measured up to 0.88. Therefore, verb constructions can be effectively used in genre classification.
Keywords: genre classification; syntax; verb construction; machine learning

1 Introduction

The existing methods of genre discrimination have good performance in classifying texts which belong to different registers (Snyman, Van Huyssteen, Daelemans 2011). Among the methods of genre identification, feature engineering for machine learning is currently popular and such methods as Naive Bayes classification, Support Vector Machine, Decision Trees and 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 fiction texts part-of-speech (POS) (Karlgren, Cutting 1994) characteristics are used; morphological, syntactical and lexical 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 (Stamatatos 2000).

The particular task is a classification of closely related genres. Traditional methods (based on statistical (Radošević, Dobša, Mladenić 2006) and discursive (Webber 2009) characteristics, POS-histograms (Kessler, Nunberg, Schutze 1997), etc., to the best of our knowledge, are not used to separate the closely related genres (Borisov, Osminkin 2013), at least with regard to the genre discrimination within Russian prose. This issue can potentially be solved by using not only the morphological features of the text, but also syntactical ones, such as verbal constructions.

In our study, we investigate the relationship between the argument-predicate structure and text genre. We compare the behavior of the verb and verb constructions across four genres of

popular prose: detective stories, fantasy, love stories, and science fiction, which are quite similar to each other and pose a challenge for existing classical methods of defining the text genre. The simple "bag-of-word" and TF-IDF methods cannot separate such genres, however, we hypothesize that the usage of additional features can improve the discrimination performance.

2 Dataset Description

For this study we used the data collected by Bogdan Evstratenko (2017). The data partly covers fiction teats available of the Russian segment of the Internet. We extracted four subcorpora: love stories, detectives, science fiction and fantasy (referred hereinafter as LS, DS, SF and FT, respectively). Texts were selected using authors' genre markers in metatags.

Genre	Size of corpus	Number of unique constructions	Coefficient of normalization
LS	7,400,231	130,435	1.1
DS	14,313,177	203,748	2.3
SF	16,228,321	195,612	2.6
FT	6,245,659	143,725	1

Table 1: The size and verb construction number of studied corpora.

The general structure of the data shows the good similarity with respect to the relative amount of parts of speech and the length of words and sentences (Table 2).

Nevertheless, the volume of both raw data and extracted verb constructions (see below) varied significantly between corpora. To address this bias, we introduced a coefficient of normalization, which was calculated as follows: the sizes of all corpora were divided by the size of smallest corpus, FT. Then, for all corpora all frequencies were divided by corresponding coefficient of normalization. As can be seen in Table 1 the SF corpus has less unique constructions than the DS corpus of comparable size, and LS and FT corpora behave the same way.

Genre	% noun	% verb	% adjective	Average length of word	Average length of phrase
LS	26	19	6	5	14
DS	25	19	6	5	13
SF	25	20	6	5	14
FT	25	19	6	7	15

Table 2: The general characteristics of corpora.

3 Preprocessing

The raw data were in FB2 format, which includes the text annotation for e-books. All tags belonged to annotation were deleted using a custom script. Then the files were processed with

UDPipe (Straka, Haji, Strakov 2016) implemented in R[1], which allows to create corpus with syntactic dependencies annotation. The UDPipe tool creates labels, morphological features, and a list of links with roots and dependent ones for each token (a token is defined as a complex of characters from a space to a space or a punctuation mark) in each sentence.

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

Nsubj (subject) – Ты зря прохаживаешься… (**You** shouldn't be strutting about here)

Obj (object) – Даже как вас зовут, и то не знаю! (I don't even know **your** name!)

Ccomp (clausal complement) – Да еще его убеждала, что ей нельзя делать аборт… (She even tried to convince him **that she can't get an abortion**)

Xcomp (open clausal complement) – Хотел ее встретить и заблудился. ([мест] wanted **to meet** her there and got lost)

Obl (oblique nominal) – Ляпнул Леня и тут же пожалел об этом. (Lenya blurted it out and immediately regretted **it**)

Advcl (adverbial clause modifier) – И даже попугай сегодня не показывался, хоть и обожал скандалы. (And even the parrot didn't show up today, **even though he loved it when somebody made a scene**).

We take into account both types of dependencies and their order in the sentence. This approach allows for estimation of core and peripheral constructions (e.i. constriction with agents); the latter are significantly undervalued in practical studies. The order of elements in the sentence increases the number of construction combinations and takes into account inversion 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.

4 Features used for machine learning

The one of the main methods of genre classification is a machine learning (Ikonomakis, Kotsiantis, Tampakas 2005), (Jindal, Malhotra, Jain 2015), (Bujlova 2018). We used several

1 https://github.com/bnosac/udpipe

sets of features for the machine learning in our study: the simplest one (the length of the sentence, word and the percentage of different parts of speech) to build the baseline model, and then more complicated ones involving deep linguistics knowledge (verb constructions). For the higher resolution of the method we selected the most prevalent constructions for LS, DS, SF and FT corpora using raw frequencies and log-likelihood (a threshold of 20 was used for frequencies and 50 for log-likelihood).

The logic behind the use of verbal constructions is based on the fact that there are three types of verbs in our corpora. The verbs of the first type have similar constructions in different genres - '*ломаться*' (to break), '*прикусить*' to bite; the verbs of the second type have different frequencies in different genres - *'познавать'* to cognize, *'поехать'* to go; 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. We created the lists of marker verbs for each corpus and applied it as a feature (the selected examples are represented in Table 3).

Verbs of the third type were defined as follows: for each verb in each genre, three most frequent constructions were singled out, then the constructions of the same verb were compared. If the constructions did not coincide in order of decreasing frequency, the verb was assigned a marker value.

	LS	DS	SF	FT
Лукавить *to cunning*	nsubj; obl nsubj; advcl	nsubj nsubj; advcl obl	nsubj nsubj; obl obl; nsubj	nsubj; obl nsubj; advcl
Бороздить *to furrow*	nsubj;obl nsubj;obj nsubj	nsubj;obj obj nsubj	nsubj;obj nsubj;obl obl	nsubj;obl;obj obl obl;obj
Расследовать *to investigate*	obj nsubj;obl nsubj	nsubj nsubj;obj obl; nsubj	obj nsubj;obl obl	obj nsubj nsubj;obl

Table 3: Selected examples of the marker verbs with top constructions.

We also developed an additional metrics – a construction of a specific cumulative threshold (CSCT), which determines the number of different constructions attached to the verb which cover N% of occurrences in a particular genre. We applied a threshold of 50% in our study. The verbs were classified depending the number of constructions that covering 50% occurrences, the verbs were defined as first-ranked, second-ranked and so on. The analyses show that in LS corpus the frequencies of first- and second-ranked verbs (189 234 and 177 983, respectively) are almost equal, and third- and fourth-ranked verbs (100 387 and 42 873, respectively) are less frequent. The rank distribution is quite similar for DE (370 714, 368 681), SF and FT.

5 Machine learning experiments

For the genre classification, we applied three common methods of machine learning: Naïve Bayes, Decision Trees and Random Forest. The simplest text features (average length of phrase, average length of word and percentage of part of speech, Table 1) were used for the definition of baseline. We conducted two experiments: both simple and calculated by log-

likelihood and using the verb construction frequencies with CSCT. Five-fold cross-validation was used for testing. The results are shown in Table 4: the use of syntactic features shows limited gain on Naïve Bayes but significantly improves the results of a Random Forest.

	Naïve Bayes			Decision Trees			Random Forest		
	precision	recall	f-score	precision	recall	f-score	precision	recall	f-score
Baseline set	0.43	0.47	0.43	0.47	0.47	0.46	0.50	0.53	0.51
Baseline+ LL-Score set	0.60	0.58	0.57	0.73	0.80	0.79	0.82	0.85	0.83
Syntax set	**0.66**	**0.66**	**0.66**	**0.86**	**0.85**	**0.85**	**0.88**	**0.88**	**0.88**

Table 4: The results of machine learning.

6 Conclusion

In our study, we analyzed four fiction corpora: love stories, detectives, science fiction and fantasy. The general structure of corpora with respect to the ratio of parts of speech and average length of sentence and word (Table 2) was highly similar in each corpus. However, as shown in Table 1, the number of verb constructions varied between corpora and was not dependent on corpus size. For instance, in the SF corpus the number of unique constructions is less than in DS corpus. This fact may be an indirect evidence of the simpler syntax in detective stories. It let us assume that syntactical features can help distinguish genres, therefore, our study addresses the application of verb constructions as features of machine learning.

After preprocessing we obtained the lists of constructions and their frequencies for each corpus. We identified three types of verbs which can possibly make a different contribution in the machine learning models: first of them were identical in all genres, second group had the same constructions across all corpora but varied in frequencies between them and the third group had different types of constructions in different corpora. The latest type was the most interesting one, covering the verbs with qualitative difference between corpora: for them different constructions were commonly used (Table 3). Some of these verbs are both syntactic and lexical markers, for example the verb *бороздить* (to furrow) has *nsubj;obl* in LS and *nsubj;obl;obj* in FT.

We introduced a new syntactic property, a construction of a specific cumulative threshold (CSCT) reflecting the number of most used constructions of the verb. The majority of verbs had 1 to 4 constructions covering the half of verb occurrences. 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 root verbs. Despite the similar distribution of the ranks in all genres, there were a number of important differences for individual verbs. Firstly, the rank of the verb is very stable across genres. Secondly, the top three constructions of the verb were often the same but appeared in different order.

We aimed to evaluate the contribution of CSCT into training, so the models of the machine learning were intentionally used with the default settings. Our baseline set (Naïve Bayes based on the length of the sentence and the percentage of different parts of speech) results in 0.43 precision and 0.47 recall. The applying of more complex approaches (Decision Tree and

Random Forest) gave 0.73 precision and 0.88 recall, respectively. Introducing of syntactical features allowed to improve machine learning results up to 0.88 precision and 0.88 recall. The data is noisy, so the results require detailed analysis - for example, the selective manual inspection of more complex constructs (having five or more dependencies) showed that their origin is more likely to be associated with machine annotation errors than with unique features of head verbs. However, such errors may occur in all buildings, which reduces their significance.

To sum up, our results showed that application of syntactical annotation corpora can be exploited for the efficient classification of closely related genres. A future expansion of current study will be tuning the machine learning models and creating the easily accessible a user-friendly pipeline.

7 References

Borisov, L.A., Orlov, Ju. N., Osminin, K.P. (2013). Identifikacija avtora teksta po raspredeleniju chastot bukvosochetanij. Prikladnaja informatika. [in Russian: Identification of a text author by the letter frequency empirical distribution]. 26(2), 95-108.

Bujlova, N. (2018). Klassifikatcia textov po zhanram pri pomoshci algoritmov mashinnogo obuchenia [in Russian: Genre classification using machine learning algorithms]. In NTI, 2(8).

Evstratenko, B. (2016). Avtomaticheskoe raspoznavanie semantycheskih poley s privlecheniem bolshih dannyh. [in Russian: Automatic recognition of semantic fields with the involvement of big data]. Barchelor thesis. NRU HSE, Moscow, Russia.

Ikonomakis, M., Kotsiantis, S., and Tampakas, V. (2005). Text classification using machine learning techniques. In WSEAS Transactions on Computers, 4(8).

Jindal, R., Malhotra, R., Jain, A., (2015). Techniques for text classification: Literature review and current trends. Accessed at: http://www.webology.org/2015/v12n2/a139.pdf [07/01/2018].

Karlgren, J., Cutting, D. (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In Proceedings of the 15th Conference on Computational Linguistics, Volume 2. pp 1071–1075, Stroudsburg, PA, USA.

Radošević, D., Dobša, J., Mladenić, D., Novak, M., Stapić, Z. (2006). Genre Document Classification Using Flexible Length Phrases. In Information and Intelligent Systems, Fakultet organizacije i informatike, Varaždin.

Snyman, D.P., Van Huyssteen, G.B., Daelemans, W. (2011). Automatic Genre Classification for Resource Scarce Languages. In Proceedings of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa, pp. 132– 137.

Stamatatos, E., Fakotakis, N., Kokkinakis G. (2000). Automatic text categorization in terms of genre and author. In Computational linguistics, 26(4), pp. 471–495.

Straka, M., Haji J., Strakov, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In LREC, pp. 4290-4297.

Webber, B. (2009). Genre distinctions for Discourse in the Penn TreeBank. Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pp. 674–682.