Text Classification Based on Deep Textual Parsing

Boris Galitsky[1], Dmitry Ilvovsky[2]

[1] Knowledge Trail Inc., San Jose, USA
bgalitsky@hotmail.com
[2]National Research University Higher School of Economics, Moscow, Russia
dilvovsky@hse.ru, echernyak@hse.ru, skuznetsov@hse.ru

Abstract. The problem of classifying text based on the deep parsing structure is addressed. An algorithm for document classification tasks where counts of words or n-grams is insufficient is proposed. The parse tree kernel method at the level of paragraphs, based on anaphora, rhetoric structure relations and communicative actions linking phrases in the parse thicket is considered.

Keywords: text genre, style classification, rhetoric structure, discourse, meta-language

1 Introduction

The problem of genre classification (also referred as automatic genre identification, AGI) has received so far some attention of the researches. Mainly there are two tied directions of these studies:

1. To develop intelligible genre system and to collect a corpus which would represent the established genre system. Usually the texts are collected from the Web [8, 11].
2. To develop a machine classifier for classifying texts of different genres [9-14].

In this paper we will consider both style and genre classification, without paying a lot of attention on the difference between these notions. Following [1] we will refer to “style” as to specific usage of language, and to “genre” as to the category of a text, which represent its intention and aim.

There are several applications of genre classification:

1. Evaluating how many different texts are there on the Web. This application can be treated as developing a socio- or psycho-metric tool [8,11,12,13].
2. Using genre classification for improving user-based information retrieval: based on the query, the search system should provide documents of appropriate genre (for example, if the query sounds scientific enough, return scholar papers, if the query is less formal – blogs, social media) [9].
3. Recognizing document type for a document management system (like security, document recommendation, and other applications) [28].

Besides there are different attempts to genre classifications the majority of researches agree upon the following idea: the less complicated text elements are used as the features for classification, the better the results are. For example, [14,29] suggests using character n-grams to perform genre classification on Brown corpus, BNC, HGC and other corpora. In [12] the syntactic patterns, morphological patterns and character n-grams are used to build feature sets and are compared to each other. The latter allowed us to achieve the highest F-measure, while the former provides poor results. The morphological pattern based classifier does not outperform the character-based one. In [13] common words are used to form feature sets.

To perform text classification in the described domains, we employ discourse information such as anaphora, rhetoric structure, entity synonymy. Relying on syntactic parse trees would provide us with specific expressions and phrasings connected with a style of writing. However, it will still be insufficient for a thorough description of linguistic features inherent to a style of writing. It is hard to identify such features without employing a discourse structure of a document. This discourse structure needs to include anaphora and rhetoric relations. Furthermore, to systematically learn these discourse features associated with the style of writing one needs a unified approach to classify graph structures at the level of paragraphs [16].

The design of such features for automated learning of syntactic and discourse structures for classification is still done manually today. To overcome this problem, tree kernel approach has been proposed [26]. Tree kernels constructed over syntactic parse trees, as well as discourse trees [17] is one of the solutions to conduct feature engineering. Convolution tree kernel [24] defines a feature space consisting of all subtree types of parse trees and counts the number of common subtrees to express the respective distance in the feature space.

The kernel ability to generate large feature sets is useful to assure we have enough linguistic features to differentiate between the classes, to quickly model new and not well understood linguistic phenomena in learning machines. However, it is often possible to manually design features for linear kernels that produce high accuracy and fast computation time whereas the complexity of tree kernels may prevent their application in real scenarios. SVM [19] can work directly with kernels by replacing the dot product with a particular kernel function. This useful property of kernel methods, that implicitly calculates the dot product in a high-dimensional space over the original representations of objects such as sentences, has made kernel methods an effective solution to modeling structured linguistic objects [25].

In this paper we will try to show how using more complicated and extensive syntactical information allows improving the result of genre classification. The goal of this research is to apply the learning based on high-level linguistic features for the style and genre classification task and also to estimate the influence of the corpus annotation quality to the quality of the performance.

2 From style to genre

Moving from “simple” to “complex” system of style classes we start to distinguish texts between 2 classes: description (object-level) and meta-description (meta-language or meta-level). We consider domain of literature documents.

A combination of object-language and metalanguage patterns and description styles can be found in literature. In the literature domain, we attempt to draw a boundary between the pure metalanguage (works of literature with a special level of abstraction) and a mixed level text (a typical work of literature). Describing the nature, a historical event, an encounter between people, an author uses a language object. Describing the thought, beliefs, desires and knowledge of characters about the nature, events and interactions between people, an author may use a metalanguage, if its entities/ range over the expressions (phrases) of the language-object.

An outstanding example of the use of metalanguage in literature is Franz Kafka's novel "The Trial". According to our model, the whole plot is described in metalanguage, and object-level layer is not presented at all. This is unlike a typical work of literature, where both levels are employed and object-level prevail, such as fairy tales. The novel is a pure example of the presence of meta-theory and absence of object-level theory, from the standpoint of logic. The reader is expected to form the object–level theory herself to avoid an ambiguity in the interpretation of this novel.

For the genre classification we used the system of genres and the corpus from [2,3]. Let us describe the genre system in more details. Unlike in other disciplines, authors do not define particular genres in systematic, exact, crisp way, but instead of this construct 17 main so-called Functional Dimension which are the basis for a genre description. For example, the direction A7 corresponds to instructions (tutorials, FAQs, manuals, recipes), the direction A11 – to personal writing, such as diary-like blogs, personal letters, traditional diaries. A collection of texts, picked from the Web, is annotated by humans according to these directions: the annotator is asked to what extent this or that direction is present in the text. There are four possible answers: 0 none or hardly at all; 0.5 slightly; 1 somewhat or partly; 2 strongly or very much so. After the annotation, every text is represented as a vector in the space of 17 functional dimensions, which makes any kind of machine learning applicable. The texts and functional dimension are bi-clustered and the resulting clusters are said to represent a genre. The resulting system of genres consists of combinations of FTDs. Let us describe some of genres, achieved in [2,3]. There are genres that use only singly dimension: for example, the cluster Cl6 corresponds to the dimension A16, which is aimed at presenting information. But the are some genres that correspond to two or three dimensions: the cluster Cl13 stands for dimensions A1 + A11, which are opinion blogs, often reporting personal experience and expressing one's emotions; and the cluster Cl14 stands for dimensions A11 + A19 + A3, which are diary blogs expressing one's emotions and attempting to embellish the description. The clusters often correspond to traditional genres, but are more reliable than traditional genres, since the annotator does not have to choose between several predefined genres. We adopt both the genre system and the corpus from this research.

3 Discourse text structure for the classification task

It turns out that low-level features could be insufficient for the style classification in some domains like meta-document or design-document text detection. Since important phrases can be distributed through different sentences, one needs a sentence boundary – independent way of extracting both syntactic and discourse features. Therefore we intend to combine/merge parse trees to make sure we cover all the phrase of interest.

Rhetorical Structure Theory (RST) [5, 20] has been used to describe or understand the structure of texts and to link rhetorical structure to other phenomena, such as anaphora or cohesion. RST is one of the most popular approach to model extra-sentence as well as intra-sentence discourse. RST represents texts by labeled hierarchical structures. Their leaves correspond to contiguous Elementary Discourse Units; adjacent ones are connected by rhetorical relations (e.g., Elaboration, Contrast), forming larger discourse units (represented by internal nodes), which in turn are also subject to this relation linking. Discourse units linked by a rhetorical relation are further distinguished based on their relative importance in the text: nucleus being the central part, whereas satellite being the peripheral one. Discourse analysis in RST involves two subtasks: discourse segmentation is the task of identifying the EDUs, and discourse parsing is the task of linking the discourse units into a labeled tree.

Discourse analysis explores how meanings can be built up in a communicative process, which varies between a text metalanguage and a text language-object. Each part of a text has a specific role in conveying the overall message of a given text.

4 Learning on extended parse trees

The design of discourse and syntactic features for automated text assessment tasks is still an art nowadays. One of the solutions to systematically treat these features is the set of tree kernels built over syntactic parse trees, extended by discourse relations. Convolution tree kernel [24, 25] defines a feature space consisting of all subtree types of parse trees and counts the number of common subtrees as the syntactic similarity between two parse trees. They have found a lot of applications in a number of NLP tasks.

To obtain the inter-sentence links, we employed anaphoric relations from Stanford NLP [22, 23]. Rhetoric parser [15] builds a discourse parse tree by applying an optimal parsing algorithm to probabilities obtained from two conditional random fields, intra-sentence and multi-sentence parsing. We also rely on additional tags to extend SVM feature space, finding similarities between trees. These additional tags include noun entities from Stanford NLP such as organization and title, and verb types from VerbNet.

For every arc which connects two parse trees, we obtain the extension of these trees, extending branches according to the arc. For a given parse tree, we will obtain a set of its extension, so the elements of kernel will be computed for many extensions, instead of just a single tree [17]. The problem here is that we need to find common sub-trees for a much higher number of trees than the number of sentences in text, however by subsumption (sub-tree relation) the number of common sub-trees will be substantially reduced. The resultant trees are not the proper parse trees for a sentence, but nevertheless form an adequate feature space for tree kernel learning.

5 Evaluation

5.1 Style dataset

For the literature domain, we collected 160 paragraphs as meta-documents from Kafka's novel "The Trial" as well as his other novels so that these paragraphs are read as metalanguage patterns. As a set of object-level documents we manually selected 200 paragraphs of text in the same domain (scholarly articles about "The Trial"). We split the data into 3 subsets for training/evaluation portions and cross-validation.

Table 1. Evaluation results for literature documents

Method	Precision	Recall	F-measure
Nearest neighbor classifier (TF*IDF based)	48.5	54.3	51.24
Tree kernel – regular parse trees	63.3	68.7	65.89
Tree kernel SVM – extended trees for both anaphora and RST	**71.5**	**73.1**	**72.29**

Table 1 shows evaluation results. Baseline approaches show rather low performance. The one of the tree kernel based methods improves as the sources of linguistic properties are expanded. For both domains, there is an improvement by a few percent due to the rhetoric relations compared with the baseline tree kernel SVM which employs parse trees only. The best accuracy is lower than 85%. This can be explained by a few reasons. Meta-documents can contain object-level text, such as design examples. Object level documents (genuine action-plan docs) can contain some author reflections on the writing process or direct citations (which are written in metalanguage). Hence the boundary between classes does not strictly separates metalanguage and language object. So for the better performance we need better annotated dataset.

5.2 Genre dataset

As it was mentioned earlier we adopted the genre system and the corpora from [1, 3]. The genre system is constructed in the following way. First, the Functional Text Dimensions (FTD) are defined. The FTD are genre annotations which reflect judgements as to what extent a text can be interpreted as belonging to a generalized functional category. A genre is a combination of several FTD. In other words, the genre is a point in the space, defined by FTD.

The corpus was annotated by humans. Every user was asked to evaluate texts of FTD on a scale: 0 none or hardly at all; 0.5 slightly; 1 somewhat or partly; 2 strongly or very much so. See an example of FTD and annotated texts below.

Table 2. Functional Text Dimensions

A1	Argum	To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view? ('Strongly', for argumentative blogs, editorials or opinion pieces)
A4	Fictive	To what extent is the text's content fictional? ('None' if you judge it to be factual/informative.)
A7	Instruct	To what extent does the text aim at teaching the reader how something works? (For example, a tutorial or an FAQ)
A8	Hardnews	To what extent does the text appear to be an informative report of events recent at the time of writing?
A9	Legal	To what extent does the text lay down a contract or specify a set of regulations?

In [2,3] 17 general dimensions are defined. Among them ten A1, A3, A4, A5, A6, A7, A8, A9, A11 form 7 different genres. See the explanations of these genres bellow. For further classification we will exploit these genres.

- [tells] Instructions for how to use software.
- [tele] Instructions for how to use hardware.
- [ted] Emotional speech on a political topic. Presentation of him/her self. Attempt to sound convincing.
- [synd] An article on a political event by a professional journalist.
- [news] A presentation of a news article in an objective, independent manner.
- [fict] Novels, stories, verses.
- [un] UN reports.

Table 3. Main genres used for the evaluation

Genre example	**A1**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A11**
ted/eva_zeisel_on_the_playful _search_for_beauty	1	0	0	1	0	0	0	0	2
FictDostoyev-skyF_CrimePun_II2_EN.txt	0	1	2	1	0.5	0	0	0	1
NewsGoal-com_MessiTop50_EN.txt	0.5	1	0	1	1	0	2	0	0.5
syndicate/exchange-rate-disorder	2	0	0	0	0	0.5	0.5	0	0
un/A_AC252_L13	1	0	0	0	0	0.5	0	2	0
TeleHTC_Manual_12_EN.txt	0	0	0	0	0	2	0	0	0
Tels-Goog_Answer_2feb_EN.txt	0	0	0	0	0	0.5	0	0	1

Table 4. Pairwise classification results

Classes	VCDim	Recall	Precision	#kernel evaluations	F-measure
Fict vs News	106	98.11	95.55	159841	96.81
Ted vs Synd	787	99.49	98.94	73177349	99.21
Un vs News	697	98.70	94.93	9486134	96.78
Tele vs Tells	360	96.69	90.76	1151517	93.63
Fict vs Ted	139	97.12	93.74	7557291	95.40
Fict vs Synd	192	95.21	94.23	7546911	94.72
Fict vs Un	214	94.90	95.71	4641983	95.30
Fict vs Tele	317	97.25	94.90	6547910	96.06
Fict vs Tells	301	96.51	95.61	8766391	96.06
News vs Ted	514	96.85	93.85	2619549	95.33
News vs Synd	281	97.28	96.19	7490174	96.73
News vs Tele	190	96.31	94.27	5235193	95.28
News vs Tells	231	98.28	96.15	3916727	97.20
Ted vs Un	390	96.45	97.03	5836394	96.74
Ted vs Tele	210	97.28	96.62	1612102	96.95
Ted vs Tells	187	94.52	96.06	7645104	96.81

The values of quality measures – recall, precision and F-measure – are optimistically high. The highest F-measure is achieved by classification of Ted against Synd. Both of these genres correspond to describing political topics. However the rhetorical structures for these genres are completely different. Hence we are able to learn a very efficient classifier to distinguish between these genres.

Another important point is a superior performance in the comparison with the results for the shallow-annotated dataset. Although the classes from this dataset could be roughly mapped on some genres (e.g. meta-level literature texts are corresponding with the [fict] genre) the distinction is less accurate.

6 Conclusions

We observed that using SVM TK one can differentiate between a broad range of text styles and genres. Each text style and genre has its inherent rhetoric structure which is leveraged and automatically learned. Since the correlation between text style and text vocabulary is rather low, traditional classification approaches which only take into account keyword statistics information could lack the accuracy in the complex cases.

In this paper we have presented two experiments on style and genre classifications. The style experiments were aimed at distinguishing between two types of writing and language usage: description and meta-description. These styles share the same vocabulary but the rhetoric structure of documents with descriptions and documents with meta-descriptions is fairly different.

For the genre classification part we adopted a corpus annotated with 7 different genres and conducted a series of pairwise classification between two genres. From mathematical point of view, as a part of future extension of this research we may conduct one genre against all-others-genres-together classification, which will allow us to understand how distinctive each genre is. Hence we will obtain a more complete picture of the genre system in general. If every genre is distinctive enough, it means that the whole genre system is well developed and the dimensions are adequate. However there might arise some problems because of the corpus being unbalanced: there are different numbers of texts in every genre and to tackle this problem we will have to balance the corpus.

7 References

1. Lee, David YW. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. (2001)
2. Sharoff, S. In the garden and in the jungle: Comparing genres in the BNC and Internet. In Genres on the Web: Computational Models and Empirical Studies, pages 149–166. Springer, Berlin/New York. (2010)
3. Sharoff, S., Wu, Z., and Markert, K. The Web library of Babel: evaluating genre collections. In Proc. of the Seventh Language Resources and Evaluation Conference, LREC 2010, Malta. (2010)
4. Richard Forsyth and Serge Sharoff. Document Dissimilarity within and across Languages: a Benchmarking Study. *Literary and Linguistic Computing*, 29:6–22. (2014)
5. Mann, W., Matthiessen, C., Thompson, S.: Rhetorical Structure Theory and Text Analysis. Discourse Description: Diverse linguistic analyses of a fund-raising text / ed. by W. C. Mann and S. A. Thompson. – Amsterdam. –P. 39–78 (1992)
6. Egg, M., Redeker, G.: Underspecified discourse representation. In: Anton Benz & Peter Kühnlein (eds), *Constraints in Discourse* (pp. 117-138), Amsterdam: Benjamins. (2008)
7. Taboada, M.: The Genre Structure of Bulletin Board Messages. Text Technology 13 (2): 55-82. (2004)
8. Biber, Douglas, Jerry Kurjian. Towards a taxonomy of web registers and text types: a multidimensional analysis." Language and Computers 59.1 (2006): 109-131. (2006)
9. Freund, Luanne, Charles LA Clarke, Elaine G. Toms. Towards genre classification for IR in the workplace. Proceedings of the 1st international conference on Information interaction in context. ACM (2006).
10. Kessler, Brett, Geoffrey Numberg, Hinrich Schütze. Automatic detection of text genre." Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics (1997).
11. Santini, Marina. Automatic identification of genre in web pages. Diss. University of Brighton (2007).
12. Sarawgi, Ruchita, Kailash Gajulapalli, Yejin Choi. Gender attribution: tracing stylometric evidence beyond topic and genre. Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics (2011).
13. Stamatatos, Efstathios, Nikos Fakotakis, George Kokkinakis. Text genre detection using common word frequencies." Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics (2000).

14. Wu, Zhili, Katja Markert, and Serge Sharoff. Fine-grained genre classification using structural learning algorithms. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2010).
15. Joty, S., Carenini, G., Ng, R., Mehdad, Y.: Combining Intra- and Multi-sentential Rhetorical Parsing for Document-level Discourse Analysis.In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria (2013)
16. Galitsky, B., Ilvovsky, D., Kuznetsov, S. O., Strok, F.: Matching sets of parse trees for answering multi-sentence questions // Proceedings of the Recent Advances in Natural Language Processing, RANLP 2013. – INCOMA Ltd., Shoumen, Bulgaria.– P. 285–294 (2013)
17. Ilvovsky, D.: Going beyond sentences when applying tree kernels. Proceedings of the Student Research Workshop.– ACL 2014.– P. 56-63 (2014)
18. Galitsky, B., Kuznetsov, S.O.: Learning communicative actions of conflicting human agents. J. Exp. Theor. Artif. Intell. –Vol. 20(4). – P. 277–317 (2008)
19. Vapnik, V.: The Nature of Statistical Learning Theory. – Springer-Verlag (1995)
20. Marcu, D.: From Discourse Structures to Text Summaries. Proceedings of ACL Workshop on Intelligent Scalable Text Summarization / eds. I. Mani and M. Maybury. – Madrid, P. 82–88 (1997)
21. Severyn, A., Moschitti, A.: Fast Support Vector Machines for Convolution Tree Kernels. Data Mining Knowledge Discovery 25.– 2012.– P.325-357. (1997)
22. Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. The Life and Death of Discourse Entities: Identifying Singleton Mentions. In Proceedings of NAACL 2013.
23. Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics 39(4), 2013.
24. Collins, M., Duffy, N.: Convolution kernels for natural language. In Proceedings of NIPS, 625–632 (2002)
25. Moschitti, A.: Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. In Proceedings of the 17th European Conference on Machine Learning, Berlin, Germany (2006)
26. Sharoff, Serge. "Functional Text Dimensions for annotation of Web corpora. http://corpus.leeds.ac.uk/serge/publications/2015-corpora-submission.pdf (2015)
27. Cumby, C. and Roth D. On Kernel Methods for Relational Learning. ICML, pp. 107-14. (2003)
28. Galitsky B., Ilvovsky D., Kuznetsov S. O. Text Classification into Abstract Classes Based on Discourse Structure, in: Proceedings of the Recent Advances in Natural Language Processing, RANLP 2015. Hissar. P. 201-207. (2015)
29. Kanaris, I. and E. Stamatatos Learning to Recognize Webpage Genres Information Processing and Management, 45(5), pp. 499-512, Elsevier (2009)